LECTURE 3: STATISTICAL DECISION THEORY

Michal Kolesár[*]

September 9, 2024

Statistical decision theory is an organizing framework that underlies everything we do in statistics, including estimation and hypothesis testing, but also many other things such as the design of experiments, designing sequential stopping rules (should we stop the experiment and report an estimate, or else wait and collect more data?), or multiple decision problems (e.g. accept, reject, recommend for further study; another example: is the new drug, better, worse, or as good as the old drug?).

In all of these problems the job of a statistician is to come up with a *decision rule*: mapping from the data to a decision (whether to accept the test or stop the experiment, what estimate to report, etc). We first outline the basic elements of this framework, and then discuss how the framework can be used to evaluate and compare different decision rules. While statistical decision theory tells us how to *evaluate* different decision rules, it does not tell us how to *find* good ones. We'll spend much of the rest of the course discussing several common strategies, both in the context of estimation, and in the context of testing.

1. BASIC CONCEPTS

REFERENCE Background reading: Chapter 1 in Ferguson (1967).

Digression. Much of game theory originates with von Neumann and Morgenstern (1944). Abraham Wald then made the connection between their ideas and the theory of statistics, which he laid out in his landmark book (Wald 1950). The book laid the ground for statistical decision theory as we know it today. Indeed, the term "statistical decision theory" is a condensation of Wald's phrase "the theory of statistical decision functions". This theory generalized the theory of tests and confidence intervals developed by Neyman and Pearson by showing that other types of problems in statistics (estimation, optimal stopping rules, etc.) can be treated similarly. ☒

A statistical decision problem can be thought of as a two-person game: statistician against nature. This game has 3 elements: A set Θ called the *parameter space*, an *action space* $\mathcal{A}$ available to the statistician, and a loss function $L\colon \mathcal{A}\times\Theta \to \mathbb{R}_+$. Nature chooses a parameter $\theta \in \Theta$, which the statistician doesn't see (θ is sometimes called state of nature), and a statistician chooses an *action* $a \in \mathcal{A}$. The loss function $L(a, \theta)$ reflects the

[*] Email: `mkolesar@princeton.edu`.

loss (negative utility) to the statistician when they take action a and the parameter is θ. Ideally, we choose the loss function based on the economics of the problem; often, however, we choose it to qualitatively reflect what we are trying to do and to be mathematically convenient. In contrast to classic game theory, we don't make assumptions about nature's payoff function: nature, in particular, doesn't have to act rationally.

The second point of departure from game theory is that the statistician is allowed to gather some information about nature's action by observing data $X \sim F(x \mid \theta)$, where F is the cumulative distribution function (CDF) of X. A *statistical decision problem* is the triple $(\Theta, \mathcal{A}, L)$ coupled with the data X. We denote the probability density function (PDF) of X by $f(x \mid \theta)$. For a fixed x, $f(x \mid \theta)$ as a function of θ is called the *likelihood function*. Sometimes we write F_θ rather than $F(x \mid \theta)$. The set of possible distributions $\{F_\theta \colon \theta \in \Theta\}$ is called a *statistical model*. If Θ is finite-dimensional, the model is *parametric*, otherwise it is non-parametric.

Remark 1 (Random sample). Typically, the data distribution $F(\cdot \mid \theta)$ will have a particular structure, depending on how the statistician collected it. The simplest and perhaps most common case is that the X consists of multiple independent observations, each of which has the same distribution: $X = (X_1, \dots, X_n)$, with X_i i.i.d. We refer to this as having a *random sample*. That is, $\{X_i\}_{i=1}^n$ is *a random sample* of size n from distribution $F(\cdot \mid \theta)$ if $X_1, \dots, X_n$ are mutually independent and $X_i \sim F(\cdot \mid \theta)$ for each i.

Thus, a random sample is an i.i.d. sample: the observations are independent, and they are drawn from the same distribution. If $X_1, \dots, X_n$ is a random sample from a distribution with PDF $f(x \mid \theta)$, then the joint PDF is, by independence, given by

$$f(x_1, \dots, x_n \mid \theta) = \prod_{i=1}^n f(x_i \mid \theta),$$

where we abuse notation and write $f(x_1, \dots, x_n \mid \theta)$ rather than $f_{X_1,\dots,X_n}(x_1, \dots, x_n \mid \theta)$.

This i.i.d. business may seem a bit mysterious at this point. You can choose to either accept it, or else read the optional Section 4 below that provides an elegant motivation for this assumption.

The statistician's job is to pick an action $a \in \mathcal{A}$ after observing $X = x$. If for each x in the sample space $\mathcal{X}$, this decision is deterministic, we can characterize it by the (non-randomized) *decision rule* $\delta(X)$, where $\delta \colon \mathcal{X} \to \mathcal{A}$. More generally, the decision rule may be random (sometimes this is a useful technical device; at other times the ability to randomize—to play a mixed strategy—is essential. Can you think of such situations?). The statistician has to commit to a decision rule before seeing X, at which point the loss is random. Therefore, the statistician evaluates the possible rule δ according to their expected loss, or *risk*,

$$R(\delta, \theta) = E_\theta[L(\delta(X), \theta)].$$

The risk function represents the average loss when the true state of nature is θ and the statistician uses the rule $\delta(\cdot)$. The key point is that the choice of δ should depend only on the risk function.

Example 1 (Hypothesis testing). In *hypothesis testing*, we are interested in testing the null $\theta \in \Theta_0$, for some $\Theta_0 \subseteq \Theta$. Here the action space is $\mathcal{A} = \{\text{accept}, \text{reject}\}$. The loss is ℓ_I if we make a type I error (reject a true null), ℓ_{II} if we make a type II error (accept a false null), and zero otherwise. The risk is ℓ_I times the level of the test if $\theta \in \Theta_0$, and it is ℓ_{II} times the power of the test if $\theta \in \Theta_1$. The decision rule is called a *test*. We'll study this problem in detail later. ⊠

Example 2 (Estimation). In *estimation*, we are interested in directly learning about nature's move, θ, so that $\mathcal{A} = \Theta$. The decision rule is called an *estimator*, and we usually use the notation $\hat{\theta}(X)$ rather than $\delta(X)$, often dropping the argument and just writing $\hat{\theta}$. The particular action $\hat{\theta}(x)$ that we take is called an *estimate*. If Θ has many dimensions, we're only typically interested in part of it. If $\theta = (\beta, \gamma)$ and we only want to estimate β and don't care about γ, then γ is called a *nuisance parameter*, and β is called the *parameter of interest*.

There are many possible loss functions that one can use for evaluating estimators. When $\dim(\theta) = 1$, we can use the absolute value loss, $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$, zero-one loss $L(\hat{\theta}, \theta) = \mathbb{1}\{|\hat{\theta} - \theta| > k\}$ where k is some constant, etc. However, the most common loss function for evaluating estimators is quadratic loss $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, the risk of which is called mean squared error (MSE):

$$\begin{aligned}
\text{MSE}(\hat{\theta}, \theta) &= E[(\hat{\theta} - \theta)^2] \\
&= E[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)^2] = \text{var}(\hat{\theta}) + (E[\hat{\theta}] - \theta)^2 + 2E[\hat{\theta} - E[\hat{\theta}]](E[\hat{\theta}] - \theta) \\
&= (E[\hat{\theta}] - \theta)^2 + \text{var}(\hat{\theta})
\end{aligned}$$

The quantity $E[\hat{\theta}] - \theta$ is called the *bias* of the estimator, and therefore we have the MSE decomposition

$$\text{MSE} = \text{bias}^2 + \text{variance}.$$

Both the bias and the variance in general depend on θ. If the bias is zero for all θ, then the estimator is called *unbiased*. ⊠

Remark 2 (Risk). We said that the choice of δ should depend only on the risk function. This raises two questions: why should we use expected loss, and how to interpret the probability P_θ (induced by $F(\cdot \mid \theta)$) used to calculate the expectation. One interpretation is that these probabilities are "objective", or "common knowledge", in which case we motivate using risk in the same way we motivate using expected utility in microeconomic theory: by the expected utility theorem (Mas-Collel, Whinston, and Green 1995, Proposition 6.B.3). Alternatively, the distribution of X may not be known, even if we knew θ. However, if the statistician has well-behaved preferences over the possible rule δ if θ were known, then this is there exists a probability P_θ such that the statistician prefers δ to δ' iff $R(\delta, \theta) \leq R(\delta', \theta)$. This is called the *subjective expected utility theorem* (see Mas-Collel, Whinston, and Green (1995, Proposition 6.F.1) or Ferguson (1967, Section 1.4)). We can interpret P_θ as the subjective belief of the statistician about the behavior of X.

Remark 3. The reason to use von Neumann-Morgenstern decision theory is that it ensures coherence of our decisions, much like probability axioms ensure coherence when assigning probabilities. You can start out by specifying a loss and probabilities, and deduce decision recommendations, or you can go the other way around, deduce the loss function and the probabilities from decisions. What's more, you can go back and forth, adjusting the loss, the probabilities, or the decisions until you iron out all the inconsistencies and incoherence.

Ideally, we would like to construct a decision rule that is at least as good as all other decision rules for all θ, i.e. pick a rule $\delta(X)$ such that $R(\delta,\theta) \leq R(\delta',\theta)$ for all rules δ'. That is generally impossible. However, it's reasonable to impose a much weaker requirement of *admissibility*.

Definition 4. An decision rule δ is called *inadmissible* if there exists another rule δ^* with uniformly smaller risk, $R(\delta^*,\theta) \leq R(\delta,\theta)$ for all θ, with the inequality being strict for at least one θ. A decision rule that is not inadmissible is *admissible*.

Note that admissibility is defined with respect to a specific loss function, although for certain problems there exist decision rules that are admissible for a large class of loss functions.

Example 3. Suppose that $X = (X_1, \ldots, X_n)$ with $X_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d, and σ^2 known, so that $\theta = \mu$, and $\Theta = \mathbb{R}$. Then $\overline{X}_n$ is unbiased for μ, and has variance σ^2/n. Therefore,

$$\text{MSE}(\overline{X}_n, \mu) = \sigma^2/n.$$

Now consider estimating μ using the sample median, $X_{\lceil n/2 \rceil}$. It's also unbiased (by symmetry). Laplace showed that the variance of the sample median is approximately (for n large) equal to $\frac{1}{4nf(m)^2}$, where f is the PDF of X_i and m is the median of X. In our case, this implies

$$\text{var}(X_{\lceil n/2 \rceil}) \approx \frac{1}{4n\phi(0)^2} = \frac{2\pi\sigma^2}{4n} = \frac{\pi}{2}\frac{\sigma^2}{n} > \frac{\sigma^2}{n}.$$

Therefore, the sample median is not admissible (under quadratic loss).

On the other hand, the estimator $\hat{\mu} = 1$ (that disregards the data) is admissible, since no other estimator achieves exactly zero risk at $\mu = 1$. ⊠

Admissibility is a very weak requirement: for a typical estimation or testing problem, there will be a lot of admissible estimators or tests.

This raises the question of how to choose the "best" decision rule among the admissible ones. One option is to further restrict the class of decision rules. The two most common ways of doing this is by requiring unbiasedness (discussed in Section 2 below), or appealing to the invariance principle (which we won't discuss). An alternative to restricting the class of decision rules is to turn the partial order implied by the risk function into a complete ordering. The two most common ways of doing this are

1. Use some weight function $\pi(\theta)$ (that is non-negative and integrates to one over Θ) and rank the decision rules by their weighted average risk, called the *Bayes risk*

$$r(\delta, \pi) = \int_{\Theta} R(\delta, \theta)\pi(\theta)\, \mathrm{d}\theta.$$

   An decision rule that minimizes this risk is called *Bayes rule* (with respect to the weighting function $\pi(\theta)$).

2. Use the worst-case risk:

$$\overline{R}(\delta) = \sup_{\theta \in \Theta} R(\delta, \theta).$$

   A decision rule that minimizes the worst-case risk is called *minimax*. This criterion was proposed by Wald (1950).[1]

There are two justifications for using Bayes risk, analogous to the two justifications for using expected loss in Remark 2. First, we may know that nature chooses θ according to the density π (maybe θ is the outcome of a coin toss), and all we are doing is taking the expectation over θ as well as X. The other justification is that π reflects the subjective beliefs about θ by the statistician; they are called the statisticians' *prior*. By the subjective expected utility theorem, such a prior distribution exists provided that the preference patterns of the statistician are reasonable. The beauty of this approach is that, as we'll see in a future lecture when we discuss Bayes rules in detail, all the statistician has to do is to use the data to update their beliefs, forming a *posterior*. Then, whatever the decision problem, it is easy to calculate the Bayes rule.

One of the few things that Fisher and Neyman agreed upon was that they argued that the statistician has seldom enough information about the unknown state of nature θ to build a prior distribution for it: effectively, they argued that the statistician's preferences are not well-defined (e.g. the independence axiom fails). Note, however, the strange asymmetry: the statistician has enough information to form beliefs about $X \mid \theta$, but somehow not enough information to form beliefs about θ. A more convincing argument against Bayes rules that is also an argument in favor of minimax rules is due to Savage (1972, Chapter 10)[2]. In particular, the statistician's goal is not just to make a decision that they believe is the best, but to communicate to an audience, or to act on behalf of a group of people, who may have different beliefs (imagine an FDA jury deciding whether to approve a drug, or communicating your research findings to the research community). This group agrees on the likelihood $f(x \mid \theta)$, but each member may have different beliefs $\pi(\theta)$ (we can always make sure that the audience agrees on the likelihood by making the model sufficiently rich: in fact, that's what we do in research!). To make sure everybody in the group agrees on a common decision rule δ, we pick a rule that ensures that the expected loss, as perceived by each member, is acceptable: we pick a rule that minimizes

[1] Wald did not offer much in a way of motivation, writing in Chapter 1: "a minimax solution seems, in general, to be a reasonable solution of the decision problem when an a priori distribution for θ does not exist or is unknown to the experimenter".

[2] Savage later said that minimax rules are "ill-founded".

the maximum risk,

$$\sup_{\pi \in \mathcal{P}} r(\delta, \pi) = \overline{R}(\delta),$$

where $\mathcal{P}$ is the set of priors held by the audience, and the equality holds if $\mathcal{P}$ is sufficiently rich. To see this, note that if $\mathcal{P}$ contains mass points (degenerate priors that put all probability on one θ), then clearly $\sup_{\theta \in \Theta} R(\delta, \theta) \leq \sup_{\pi \in \mathcal{P}} R(\delta, \theta)$. In the other direction, for any π, $\int R(\delta, \theta)\pi(\theta)d\theta \leq \int \sup_{\theta \in \Theta} R(\delta, \theta)\pi(\theta)d\theta = \overline{R}(\delta)$. This justifies the use of a minimax rule if our goal is to act on behalf of or to communicate to a diverse audience. Note this justification can also be used for the requirement we imposed above that we pick the rule δ before we see the data—otherwise the audience may worry about p-hacking etc.

Digression. In contrast, defending the minimax principle on purely frequentist grounds appears to be difficult, as argued in Savage (1972, Chapter 9.7). Suppose we have to decide whether to bet $2 for or against the event that electric cars will entirely replace the internal combustion engine next year. The minimax solution is to bet for with probability 1/2, and bet against probability 1/2: few people would find this reasonable. ☒

Example 4. Have you ever disagreed with a friend about which restaurant to go to, and flipped a coin to resolve the issue? Then you employed a minimax decision rule. Drawing random samples is also a minimax rule. ☒

Example 3 (continued). MSE of the sample mean doesn't depend on μ, so that both the average and maximum risk of $\overline{X}_n$ is σ^2/n. Furthermore, it can be shown that $\overline{X}_n$ is minimax and admissible.

Now consider estimating $\mu \in \mathbb{R}^k$ based on $X_i \sim \mathcal{N}(\mu, \sigma^2 I_k)$, with σ^2 known, using the compound loss $L(\hat{\mu}, \mu) = \sum_{j=1}^{k} (\mu_j - \hat{\mu}_j)^2$. You may be tempted to generalize the single-dimensional case and say that $\hat{\mu} = \overline{X}_n$ is still minimax and admissible. This turns out to only be half-right, since while $\overline{X}_n$ is minimax, it's not admissible if $k \geq 3$ (Stein 1956). In particular, it's dominated by several shrinkage estimators, such as the James and Stein (1961) estimator

$$\hat{\mu}_{JS} = \left(1 - \frac{(k-2)\sigma^2/n}{\|\overline{X}_n\|^2}\right) \overline{X}_n,$$

where $\|x\| = \sqrt{\sum_i x_i^2}$ is the Euclidean norm. The James-Stein estimator thus pulls the sample mean towards zero. This is a surprising result (what's special about zero? What's the intuition for the inadmissibility of $\overline{X}_n$?), and it generated a new branch of statistical methods called "Empirical Bayes" methods. It also underlies many of the modern machine learning techniques, as well as large parts of nonparametric statistics. ☒

2. UNBIASED ESTIMATORS

The property of unbiasedness has an intuitive appeal, but it is a bit hard to justify. One motivation is that under squared error loss, unbiasedness is equivalent[3] to the property that

$$E_\theta[(\theta' - \hat{\theta}(X))^2] \geq E_\theta[(\theta - \hat{\theta}(X))^2],$$

that is, as measured by the mean squared error, the estimator is closer to the true θ on average than to any other θ. Here I use subscript θ on the expectation to emphasize that the expectation is under the distribution $F(\cdot \mid \theta)$. In simple problems, once we restrict attention to unbiased estimators, it is possible to find one that has minimal variance—much of old statistical theory is concern with the question of finding such minimum variance unbiased estimators. The decision-theoretic justification for this is rather weak, as noted by Hodges and Lehmann (1950):

> The principles most commonly applied in the selection of a point estimate are the principles of maximum likelihood and of minimum variance unbiased estimation. Both of these principles are intuitively appealing, but neither of them can be justified very well in a systematic development of statistics

(I disagree with their assessment of maximum likelihood, for reasons we will explore later).

The two most common estimators: the *sample mean* ($\overline{X}_n = \sum_{i=1}^n X_i/n$) and the *sample variance* ($S_n^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2/(n-1)$) are both unbiased.

Lemma 5 (Casella and Berger 2002, Theorem 5.2.6). If $X_1, \ldots, X_n$ is a random sample of size n from a population distribution with mean μ and variance σ^2, then $E[\overline{X}_n] = \mu$ and $E[S_n^2] = \sigma^2$.

Proof. By linearity of expectation,

$$E[\overline{X}_n] = E[\sum_{i=1}^n X_i/n] = \sum_{i=1}^n E[X_i]/n = \sum_{i=1}^n \mu/n = \mu.$$

To show the second part of the lemma, denote $Y_i = X_i - \mu$ and $\bar{Y}_n = \sum_{i=1}^n Y_i/n$. Note that $E[Y_i] = 0$. Thus, $E[Y_i^2] = \text{var}(Y_i) = \text{var}(X_i) = \sigma^2$. Since Y_i are independent, $\text{var}(\bar{Y}_n) = \sigma^2/n$.

[3] under mild regularity conditions, see Problem 1.2 in Lehmann and Romano (2005)

Then

$$\begin{aligned}
E[S_n^2] &= E\left[\sum_{i=1}^n (X_i - \overline{X}_n)^2/(n-1)\right] \\
&= \sum_{i=1}^n E\left[\frac{((X_i-\mu)-(\overline{X}_n-\mu))^2}{n-1}\right] \\
&= \sum_{i=1}^n E\left[\frac{(Y_i-\overline{Y}_n)^2}{n-1}\right] \\
&= \sum_{i=1}^n \frac{E[Y_i^2]-2E[Y_i\overline{Y}_n]+E[\overline{Y}_n^2]}{n-1} \\
&= \sum_{i=1}^n \frac{\sigma^2-2E[Y_i\overline{Y}_n]+\sigma^2/n}{n-1} \\
&= \sum_{i=1}^n \frac{\sigma^2-2\sigma^2/n+\sigma^2/n}{n-1} \\
&= \sigma^2,
\end{aligned}$$

which completes the proof. □

While in simple settings, unbiasedness leads to intuitive estimators, sometimes unbiasedness restricts the class of estimators too much in the sense that we're left with an empty set: no unbiased estimators exist.

Example 5 (Lehmann and Casella 1998, Example 2.1.2). Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution with parameter p. Suppose that our parameter of interest $\theta = 1/p$. Let $\hat{\theta}(X)$ be some estimator. Then

$$E[\hat{\theta}(X)] = \sum_{(x_1,\ldots,x_n)\in\{0,1\}^n} \hat{\theta}(x_1,\ldots,x_n)P((X_1,\ldots,X_n)=(x_1,\ldots,x_n)).$$

We know that for any $(x_1,\ldots,x_n)\in\{0,1\}^n$,

$$P((X_1,\ldots,X_n)=(x_1,\ldots,x_n)) = p^{\sum_{i=1}^n x_i}(1-p)^{\sum_{i=1}^n(1-x_i)},$$

which is a polynomial of degree n in p. Therefore, $E[\hat{\theta}]$ is a polynomial of degree at most n in p. However, $1/p$ is not a polynomial at all. Hence, there are no unbiased estimators in this case. ☒

In some other cases, unbiased estimators may be quite weird.

Example 6. Suppose there is an infinite number of independent trials, $X_1, X_2, \ldots$ from Bernoulli distribution with parameter p. Suppose that $X_i = 0$ means failure in the ith trial while $X_i = 1$ means success. Suppose that instead of observing the sequence $\{X_i\}_{i=1}^\infty$, we observe only the number of failures before the first success. Denote the number of failures by Y. Then $P(Y = y) = (1-p)^y p$. Suppose that the parameter of interest p. Let $\hat{p}(Y)$ be some estimator. Then $E[\hat{p}] = \sum_{y=0}^\infty \hat{p}(y)(1-p)^y p$. If $\hat{p}$ is unbiased for p, then $\sum_{y=0}^{\infty} \hat{p}(y)(1-p)^y p = p$. Equivalently, for all $0 \leq p < 1$,

$$\sum_{y=0}^{\infty} \hat{p}(y)(1-p)^y = 1.$$

Thus, the only unbiased estimator is $\hat{p}(0) = 1$ and $\hat{p}(y) = 0$ for all $y \geq 1$. Does it seem reasonable? ☒

Often there is a trade-off between bias and variance: in the previous example requiring unbiasedness led us to an estimator with variance that was unreasonably large. Thus, we may prefer a slightly biased estimator to an unbiased one if the former has much smaller variance in comparison to the latter one (with the trade-off measured by MSE).

3. ASYMPTOTIC PROPERTIES OF ESTIMATORS

Sometimes, the finite-sample risk of an estimator is hard to figure out, in which case we may want to pick an estimator based on its asymptotic risk, or other asymptotic properties. As a first step towards figuring out its asymptotic risk, in most problems, we start out by establishing the following two properties:

1. An estimator $\hat{\theta}$ is *consistent* for θ if $\hat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$.
2. An estimator is *asymptotically normal* with asymptotic variance Σ if there exist sequences γ_n and a_n and a matrix Σ such that $a_n(\hat{\theta} - \gamma_n) \Rightarrow \mathcal{N}(0, \Sigma)$. a_n is called the *rate of convergence* of $\hat{\theta}$, and Σ its asymptotic variance. Often, we may be able to set $\gamma_n = \theta$ and $a_n = n^{1/2}$, in which case the estimator is called *root-n* consistent.

Remark 6. Note that Σ is not necessarily equal to the limit of the finite-sample variances of $\hat{\theta}$, $\lim_{n\to\infty} \operatorname{var}(a_n(\hat{\theta} - \gamma_n))$. In fact, the limit might not even exist, but if it does, then Σ is necessarily smaller than or equal to the limit of the variances (by Fatou's lemma from measure theory): in general, if $X_n \Rightarrow X$, then $\liminf_{n\to\infty} \operatorname{var}(X_n) \geq \operatorname{var}(X)$.[4] For instance in an instrumental variables model, the first moment of the instrumental variables estimator doesn't exist for any finite n, but the estimator is asymptotically normal.

Digression (Bickel and Doksum 1977, Exercise 4.4.9, p. 150). An example in which the inequality is strict: Let $X \sim \mathcal{U}[0,1]$, and let $X_n = X$ if $X \leq 1 - 1/n$, and $X_n = a_n$ if $X > 1 - 1/n$. Then, as an exercise, show that for any sequence a_n, $X_n \Rightarrow X$, but if $a_n/\sqrt{n} \to \infty$ (e.g. $a_n = n$), then $\operatorname{var}(X_n) \to \infty$, and if $a_n/\sqrt{n} \to a$, then $\operatorname{var}(X_n) \to a^2 + 1/12$, while $\operatorname{var}(X) = 1/12$. ☒

Digression. Because of this inequality, it is also generally not a good idea to estimate the asymptotic variance via the bootstrap. More on this in 539B...

[4] See, for instance, Lemma 3 in Hahn and Liao (2021) for a careful proof. The reason one needs to be a bit careful is that while $\liminf_{n\to\infty} E[X_n^2] \geq E[X^2]$ follows directly by Fatou's lemma, the result is not immediate if we center the random variables.

Example 7. Let $X_1, \dots, X_n$ be a random sample from some distribution with mean μ and variance σ^2. Let $\hat{\mu} = \hat{\mu}_n = \overline{X}_n$ be our estimator of μ and $S_n^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2/(n-1)$ be our estimator of σ^2. By the law of large numbers (LLN), we know that $\hat{\mu} \to_p \mu$ as $n \to \infty$. In addition, if we add and subtract μ, we get

$$\begin{aligned} S_n^2 &= \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2 = \frac{1}{n-1}\sum_{i=1}^n ((X_i - \mu) + (\mu - \overline{X}_n))^2 \\ &= \frac{1}{n-1}\sum_{i=1}^n (X_i - \mu)^2 - \frac{n}{n-1}(\overline{X}_n - \mu)^2 \\ &= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2\right) - \frac{n}{n-1}(\overline{X}_n - \mu)^2. \end{aligned}$$

By LLN, the expression in parentheses converges in probability to $E[(X_i - \mu)^2] = \sigma^2$ and $\overline{X}_n - \mu = \sum_{i=1}^n (X_i - \mu)/n \xrightarrow{p} E[X_i - \mu] = 0$. By the continuous mapping theorem (CMT), $(\overline{X}_n - \mu)^2 \to_p 0$. In addition, $n/(n-1) \to_p 1$. So, by Slutsky's theorem, $S_n^2 \to_p \sigma^2$. So $\hat{\mu}$ and S_n^2 are consistent for μ and σ^2, respectively.

Are they asymptotically normal? By the central limit theorem (CLT), $\sqrt{n}(\hat{\mu} - \mu) \Rightarrow N(0, \sigma^2)$. As for S_n^2, it follows from the last line of the preceding display that

$$\sqrt{n}(S_n^2 - \sigma^2) = \frac{n}{n-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n ((X_i - \mu)^2 - \sigma^2) - \sqrt{n}(\overline{X}_n - \mu)^2\right] + \frac{\sqrt{n}}{n-1}\sigma^2.$$

By the central limit theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n ((X_i - \mu)^2 - \sigma^2) \Rightarrow \mathcal{N}(0, \tau^2),$$

with $\tau^2 = E[((X_i - \mu)^2 - \sigma^2)^2]$. Note that $\tau^2 = \mu_4 - 2\sigma^2 E[(X_i - \mu)^2] + \sigma^4 = \mu_4 - \sigma^4$ with $\mu_4 = E[(X_i - \mu)^4]$. By Slutsky's theorem and CMT,

$$\sqrt{n}(\overline{X}_n - \mu)^2 = \frac{1}{\sqrt{n}}\left(\sqrt{n}(\overline{X}_n - \mu)\right)^2 \xrightarrow{p} 0.$$

Finally, $(\sqrt{n}/(n-1))\sigma^2 \to_p 0$. So, by Slutsky's theorem again,

$$\sqrt{n}(S_n^2 - \sigma^2) \Rightarrow N(0, \tau^2).$$

At home, try to derive this result by considering the asymptotic distribution of the vector $n^{-1}\sum_i (X_i, X_i^2)'$ and applying the delta method. ⊠

Example 8. If we have a random sample $X_1, \dots, X_n$ of size n, the CDF of the distribution that puts mass $1/n$ at each data point X_i is called the *empirical CDF*, usually denoted $\hat{F}_n$. Thus, by definition,

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i \le x\},$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, i.e. the function which equals 1 if its argument is true, and 0 otherwise. In other words, $\hat{F}_n(x)$ shows the fraction of observations with a value smaller or equal than x. The empirical CDF is unbiased and consistent:

Lemma 7. If we have a random sample $X_1, \ldots, X_n$ of size n from a distribution with CDF F, then for any $x \in \mathbb{R}$, $E[\hat{F}_n(x)] = F(x)$ and $\operatorname{var}(\hat{F}_n(x)) \to 0$ as $n \to \infty$. As a consequence, $\hat{F}_n(x) \to_p F(x)$ as $n \to \infty$.

Proof. Note that $\mathbb{1}\{X_i \leq x\}$ equals 1 with probability $P(X \leq x)$ and 0 otherwise. Thus, $E[\mathbb{1}\{X_i \leq x\}] = P(X \leq x) = F(x)$. Hence, $E[\hat{F}_n(x)] = F(x)$ by linearity of expectation. In addition, $\operatorname{var}(\mathbb{1}\{X_i \leq x\}) = F(x)(1 - F(x))$ by the formula for variance of a Bernoulli random variable with parameter $p = F(x)$. Therefore,

$$\operatorname{var}(\hat{F}_n(x)) = \sum_{i=1}^{n} \operatorname{var}(\mathbb{1}\{X_i \leq x\})/n^2 = F(x)(1 - F(x))/n \to 0. \qquad \square$$

Because $E[\mathbb{1}\{X_i \leq x\}] = F(x)$, we could alternatively have used the law of large numbers, from which the remaining claims follow directly. In particular, the strong law implies that $\hat{F}_n(x) - F(x) \overset{\text{a.s.}}{\to} 0$. Actually, an even stronger result holds:

Theorem 8 (Glivenko-Cantelli). If $X_1, \ldots, X_n$ is a random sample from a distribution with CDF F, then

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \overset{a.s.}{\to} 0.$$

Proof. See, e.g., Theorem 2.4.7 in Durrett (2019). □

This is a remarkable result: notice we have assumed *nothing* about F, and yet the theorem says that in a very strong sense, the empirical CDF will be uniformly close to it in large samples. ⊠

4. RANDOM SAMPLING (OPTIONAL)

REFERENCE Background reading: Chapter 1 in Schervish (1995).

An elegant way of motivating the random sampling framework is through the concept of exchangeability.

Definition 9. A sequence of random variables $\{X_i\}_{i=1}^{\infty}$ is *exchangeable* if for any n and m, the joint distribution of $(X_n, \ldots, X_{n+m})$ remains unchanged under permutations.

This just says that the labels i that we put on the observations don't matter. We're not taking any stance on relative frequencies or independence.

Theorem 10 (Schervish 1995, Theorem 1.47). An infinite sequence $\{X_i\}_{i=1}^{\infty}$ of Bernoulli random variables is exchangeable iff there is a unique distribution $\Pi(\theta)$ on $[0, 1]$ such that

$$P(X_1, \ldots, X_n) = \int \prod_{i=1}^{n} P_\theta(x_i) d\Pi(\theta),$$

where $P(X_i = x \mid \theta) = \theta$ and $P(X_i = 0 \mid \theta) = 1 - \theta$. Furthermore, if the sequence is exchangeable, the distribution $\Pi(\theta)$ is unique, and $\overline{X}_n - \theta \overset{a.s.}{\to} 0$ (This second part is known as de Finetti's strong law of large numbers)

So an infinite sequence is exchangeable if it's conditionally i.i.d. given *something*[5]. This is one of only two ways you can generate exchangeable random variables. The only other way: a *finite* sequence is exchangeable iff they are like draws from an urn without replacement. This is a very powerful result.

What's the implication of this result? If people believe that a random sequence is exchangeable (the labels don't matter), then they all believe that there exists a random variable θ such that conditional on it, the random variables are i.i.d. Bernoulli with parameter θ. Furthermore, θ is the long-run frequency of a success. We now see how statistical models emerge: more generally, if we have exchangeable data $\{X_i\}$ (not necessarily Bernoulli), then there is a random variable θ, such that conditional on it, the variables are i.i.d., with distribution F_θ (in the general case, we can think of θ as a one-to-one function of the limiting probability measure F). Note also that Π here can be interpreted as a prior distribution (that we may or may not want to take a stance on, as per the discussion above about Bayes vs minimax rules).

So DeFinetti's representation theorem is central to motivating statistical models, even if it's never mentioned once we write one down.

Digression. Why do we need an infinite sequence in de Finetti's theorem? Suppose $P(X_1 = 0, X_2 = 1) = P(X_1 = 1, X_2 = 0) = 1/2$, while $P(X_1 = 1, X_2 = 1) = P(X_1 = 0, X_2 = 0) = 0$. These random variables are exchangeable. But the representation doesn't hold, since it would imply $0 = \int_0^1 \theta^2 d\Pi(\theta) = \int_0^1 (1-\theta)^2 d\Pi(\theta)$, which implies that Π puts mass 1 on both 0 and 1, which is impossible. Aside: But if the sequence has length N, and we look at a small fraction $n \ll N$ of it, then the theorem is "almost" true, see Diaconis and Freedman (1980). ⊠

REFERENCES

Bickel, Peter J., and Kjell A. Doksum. 1977. *Mathematical Statistics: Basic Ideas and Selected Topics.* San Francisco, CA: Holden-Day.

Casella, George, and Roger L. Berger. 2002. *Statistical Inference.* 2nd ed. Pacific Grove, CA: Duxbury/Thomson Learning.

Diaconis, P., and D. Freedman. 1980. "Finite Exchangeable Sequences." *The Annals of Probability* 8, no. 4 (August): 745–764. https://doi.org/10.1214/aop/1176994663.

Durrett, Rick. 2019. *Probability: Theory and Examples.* 5th ed. New York, NY: Cambridge University Press. https://doi.org/doi.org/10.1017/9781108591034.

[5] For general non-Bernoulli random variables the theorem says (see Theorem 1.49 Schervish 1995) that the sequence is exchangeable iff it's i.i.d. conditional on a random probability measure. Furthermore, this random probability measure is the almost sure limit of the empirical CDF $\hat{F}_n$ (see Example 8). You may wonder what we mean by a random probability measure and by saying that it converges—we'll leave those things to an advanced probability class.

Ferguson, Thomas S. 1967. *Mathematical Statistics: A Decision Theoretic Approach.* New York, NY: Academic Press. https://doi.org/10.1016/C2013-0-07705-5.

Hahn, Jinyong, and Zhipeng Liao. 2021. "Bootstrap Standard Error Estimates and Inference." *Econometrica* 89, no. 4 (July): 1963–1977. https://doi.org/10.3982/ECTA17912.

Hodges, Joseph L., and Erich L. Lehmann. 1950. "Some Problems in Minimax Point Estimation." *The Annals of Mathematical Statistics* 21, no. 2 (June): 182–197. https://doi.org/10.1214/aoms/1177729838.

James, Willard, and Charles M. Stein. 1961. "Estimation with Quadratic Loss." In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* edited by Jerzy Neyman, 1:361–379. Berkeley, CA: University of California Press.

Lehmann, Erich L., and George Casella. 1998. *Theory of Point Estimation.* 2nd ed. New York, NY: Springer. https://doi.org/doi.org/10.1007/b98854.

Lehmann, Erich Leo, and Joseph P. Romano. 2005. *Testing Statistical Hypotheses.* 3rd ed. New York, NY: Springer. https://doi.org/10.1007/0-387-27605-X.

Mas-Collel, Andreu, Michael D. Whinston, and Jerry R. Green. 1995. *Microeconomic Theory.* Oxford, UK: Oxford University Press, June.

Savage, Leonard J. 1972. *The Foundations of Statistics.* 2nd ed. New York, NY: Dover Publications.

Schervish, Mark J. 1995. *Theory of Statistics.* New York, NY: Springer. https://doi.org/10.1007/978-1-4612-4250-5.

Stein, Charles M. 1956. "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability,* edited by Jerzy Neyman, 1:197–206. Berkeley, CA: University of California Press.

von Neumann, John, and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior.* Princeton, NJ: Princeton University Press.

Wald, Abraham. 1950. *Statistical Decision Functions.* New York: John Wiley & Sons.