In this lecture, we will introduce unbiased estimators, the concepts of uniform minimum risk unbiased (UMRU) and uniform minimum variance unbiased (UMVU) estimators, and the Lehmann-Scheffe Theorem.

1 Logistics

Homework and grading

• No midterm or final, and grading on homeworks will be lenient.

• The emphasis is on learning—feel free to collaborate/discuss on Piazza, or come to office hours for explanations from course staff.

Technical conditions

• Although these are necessary for rigorous proofs, the goal is to understand the main ideas so that we will be able to solve real problems using these ideas in the future.

• From this point on, we’ll try to avoid spending time on technical conditions whenever possible.

2 Unbiased Estimation

Definition 1 (Unbiasedness). A decision rule $\delta(x)$ is unbiased if $E_{\theta}[\delta(x)] = g(\theta) \forall \theta \in \Theta$.

Note: We will argue later in the course that unbiasedness may not be the desired property in practice, but for now our goal is to find the best unbiased estimator.

2.1 Problems with unbiasedness

Example: Suppose $X \sim \text{B}(n,p)$, i.e., $P(X = k) = \binom{n}{k}p^k(1-p)^{n-k}, 0 \leq k \leq n$. If we want to estimate $p$, we can simply use $\hat{p} = \frac{X}{n}$. Then $E[\hat{p}] = p$. But what if we want to estimate $p \ln p$?

A naive solution is $g(\hat{p}) = \hat{p} \ln \hat{p}$. Is this unbiased? Jensen’s inequality and convexity of $g(p)$ tell us

$$E[g(\hat{p})] > g(E[\hat{p}]) = g(p)$$

so this is not an unbiased estimator. What’s even worse: there does not exist any unbiased estimator for $g(p) = p \ln p$ in the Binomial model!

Proof: Suppose we have decision rule $\delta(X)$, where $X \sim \text{B}(n,p)$. Let us compute $E[\delta(X)]$ and set it equal to $p \ln p$ for all the $p \in [0, 1]$.

$$E[\delta(X)] = \sum_{k=0}^{n} \delta(k)P(X = k)$$

$$= \sum_{k=0}^{n} \delta(k)\binom{n}{k}p^k(1-p)^{n-k}$$

$$= g(p)$$

$$= p \ln p, \quad \forall p \in [0, 1]$$
But this is impossible, since \( \sum_{k=0}^{n} \binom{n}{k} \) is a polynomial in \( p \) of degree \( \leq n \), while \( p \ln p \) is not a polynomial. In general, if we are trying to estimate any quantity that can’t be written as a polynomial of degree no more than \( n \), then an unbiased estimator does not exist.

### 2.2 UMRU and UMVU

**Definition 2 (U-estimable).** If there exists an unbiased estimator for \( g(\theta) \), then \( g(\theta) \) is **U-estimable**.

Unfortunately, even if \( g(\theta) \) is U-estimable, there is no guarantee that any unbiased estimators are good in any way in a decision theoretic sense. However, it is a very healthy exercise to find the uniformly best rules among unbiased estimators.

**Definition 3 (Uniform minimum risk unbiased (UMRU)).** An unbiased estimator \( \delta(X) \) of \( g(\theta) \) is the **uniform minimum risk unbiased (UMRU)** estimator if \( R(\theta, \delta) \leq R(\theta, \delta') \), \( \forall \theta \in \Theta \), \( \forall \delta' \) unbiased.

When the squared error loss is used, we show that UMRU estimators reduce to the uniform minimum variance unbiased (UMVU) estimators, which is defined by simply replacing the condition \( R(\theta, \delta) \leq R(\theta, \delta') \) by \( \text{Var}_\theta(\delta) \leq \text{Var}_\theta(\delta') \).

The key argument is the bias-variance decomposition for the squared error loss \( L(x, y) = (x - y)^2 \):

\[
R(\theta, \delta) = E[L(g(\theta), \delta(X))] \\
= E[(\delta(X) - g(\theta))^2] \\
= E[(\delta(X) - E[\delta(X)] + E[\delta(X)] - g(\theta))^2] \\
= E[(a + b)^2]
\]

where \( a = \delta(X) - E[\delta(X)] \) and \( b = E[\delta(X)] - g(\theta) \). Then we have

\[
R(\theta, \delta) = E[a^2] + E[b^2] + 2E[(\delta(X) - E[\delta(X)])(E[\delta(X)] - g(\theta))] \\
= E[a^2] + E[b^2] + 2(E[\delta(X)] - g(\theta)) \cdot E[\delta(X) - E[\delta(X)]] \\
= E[a^2] + E[b^2] \\
= \text{Var}_\theta(\delta) + (\text{Bias}_\theta(\delta))^2.
\]

As a result, when working with squared error loss, we can always decompose our risk into a variance component and a squared bias component. Since the bias is zero when we are restricted to only use unbiased estimators, all of the risk comes from the variance (i.e., risk = variance) and the UMRU estimator is also the UMVU estimator.

### 3 Lehman-Scheffe Theorem

**Theorem 1 (Lehman-Scheffe).** Assume \( T \) is complete and sufficient, and that \( h(T) \) is an unbiased estimator for \( g(\theta) \). Then \( h(T) \) is

1. the only function of \( T \) that is an unbiased estimator for \( g(\theta) \);
2. an UMRU estimator for any convex loss function;
3. a unique UMRU estimator for any strictly convex loss function;
4. a unique UMVU estimator.

**Proof**
(1) Suppose there exists some \( \tilde{h} \neq h \) s.t. \( \tilde{h}(T) \) is also an unbiased estimator for \( g(\theta) \). Then

\[
\mathbb{E} [ \tilde{h}(T) - h(T) ] = \mathbb{E} [ \tilde{h}(T) ] - \mathbb{E} [ h(T) ] \\
= g(\theta) - g(\theta) \\
= 0 \quad \forall \theta \in \Theta
\]

If we let \( f(T) = \tilde{h}(T) - h(T) \), then \( \mathbb{E} [ f(T) ] = 0 \ \forall \theta \in \Theta \). By the definition of completeness, this implies that \( f(T) = 0 \), which implies that \( \tilde{h} = h \) everywhere. This is a contradiction, so \( h(T) \) must be the only function of \( T \) that is an unbiased estimator for \( g(\theta) \).

(2) For any unbiased \( \delta(x) \), Rao-Blackwell gives us \( \eta(T) = \mathbb{E} [ \delta(x) | T ] \) which is also unbiased (Rao-Blackwellization preserves unbiasedness), and is at least as good as \( \delta(x) \) in terms of its performance in risk function. It follows from part (1) that \( \eta(T) \) is the only function of \( T \) that is an unbiased estimator for \( g(\theta) \), which implies that \( \eta(T) = h(T) \) with probability one. Hence, we have proved that \( h(T) \) is at least as good as any unbiased estimators, which by definition shows \( h(T) \) is UMRU.

(3) Uniqueness follows from the strict convexity of the loss function.

(4) The result follows from the strict convexity of the squared error loss.

\[ \square \]

**Side note: Bahadur’s Theorem**

**Definition 4** (Minimal sufficient). Statistic \( T \) is **minimal sufficient** if for any sufficient statistic \( U \), there exists a function \( q(\cdot) \) such that \( T = q(U) \).

**Theorem 2** (Bahadur). Suppose \( T \) is complete and sufficient. Then \( T \) is minimal sufficient.

By Bahadur’s theorem, if \( T_1 \) and \( T_2 \) are two complete sufficient statistics, by minimal sufficiency there exist deterministic functions \( p(\cdot) \) and \( q(\cdot) \) such that \( T_1 = p(T_2), T_2 = q(T_1) \). Hence, the complete sufficient statistic is unique up to (bimeasurable) bijection.

### 4 Examples

**4.1 Strategies for finding UMRU estimators**

A. Rao-Blackwellization: Start with any unbiased \( \delta(x) \) and compute \( \mathbb{E} [ \delta(x) | T ] \).

- Works in theory, but in practice computing the conditional expectation can be difficult.

B. Solve \( \mathbb{E}_\theta [ \delta(T) ] = g(\theta), \forall \theta \in \Theta \).

- If you can find an unbiased estimator that is a function of \( T \), then by Lehman-Scheffe it is the only one.

C. Guess \( \delta(T) \) by playing around the distribution of \( T \).

- Wait... what?

- Although this sounds unlikely to work, we will see that in practice this can often be more reasonable than trying to work through the computations necessary for strategies A and B.
4.2 Example: Bernoulli distribution

Suppose that \( X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p) \), where \( Z \sim \text{Bernoulli}(p) \) means that \( \mathbb{P}(Z = 1) = p \) and \( \mathbb{P}(Z = 0) = 1 - p \). In order to apply any of strategies A, B, or C from above, we first need \( T \), so let’s start by finding the complete and sufficient statistic. This is easy if we can write it as an exponential family distribution:

\[
\mathbb{P}(X_1, \ldots, X_n) = p^{\sum 1(X_i = 1)} (1 - p)^{\sum 1(X_i = 0)} \\
= p^{\sum X_i} (1 - p)^{n - \sum X_i} \\
= e^{(\sum X_i) \ln p + (n - \sum X_i) \ln(1 - p)} \\
= e^{(\sum X_i) \ln \frac{p}{p} + n \ln(1 - p)}
\]

We can see that \( T = \sum X_i \) is a complete sufficient statistic. Suppose that we want to estimate \( p \), then \( \mathbb{E}(\hat{p}) = p \) implies that \( \hat{p} = n^{-1} \sum_{i=1}^n X_i \) is the UMRU estimator for \( p \). However, what if we want to estimate \( p^2 \)? Let’s try using recipe A (Rao-Blackwellization). First we need to find a simple, unbiased estimator \( \delta(x) = X_1 \cdot X_2 \):

\[
\mathbb{E}[\delta(x)] = \mathbb{E}[X_1] \cdot \mathbb{E}[X_2] = p^2.
\]

Note that \( \delta(x) \in \{0, 1\} \). We have

\[
\mathbb{E}[\delta(x)|T = t] = \mathbb{P}(X_1 = 1, X_2 = 1|\sum X_i = t) \\
= \frac{\mathbb{P}(X_1 = 1, X_2 = 1, \sum X_i = t)}{\mathbb{P}(\sum X_i = t)} \\
= \frac{\mathbb{P}(X_1 = 1, X_2 = 1, \sum_{i=3}^n X_i = t - 2)}{\mathbb{P}(\sum_{i=1}^n X_i = t)} \\
= \frac{p \cdot p^{(n-2)} (1-p)^{(n-2)-(t-2)} 1(t \geq 2)}{\binom{n}{t} p^t (1-p)^{n-t}} \\
= \frac{t(t-1) 1(t \geq 2)}{n(n-1)} \\
= \frac{t(t-1)}{n(n-1)}
\]

Therefore, our UMRU estimator for \( p^2 \) is \( h(T) = \frac{T(T-1)}{n(n-1)} \). As expected, the conditional expectation does not depend on \( p \) based on the definition of sufficiency (it should only depend on \( T \)). To summarize, we start with an unbiased estimator that is not a function of \( T \), so we cannot use Lehman-Scheffe. We first use Rao-Blackwell to find another unbiased estimator that is a function of \( T \)—then all of the implications of Lehman-Scheffe apply.