

## Sufficiency

Lecture #3: Tuesday, 11 January 2005  
Lecturer: Prof. Charles Elkan  
Scribe: Jan Voung  
Reviewer: Hyun Min Kang

### 1 Outline and review

The previous lecture showed that an estimator with minimum mean squared error is desirable, but unachievable in general. We then decided to consider only unbiased estimators. Under that restriction, we discovered that an unbiased estimator of minimum MSE also has minimum variance. So, our goal now is to find minimum variance unbiased estimators (MVUEs). Before doing so, we must go through:

1. Definitions (today)
2. Lemmas
3. A central theorem.

### 2 How to summarize data

We usually observe outcomes  $x$  and assume they are drawn from a family of distributions  $\{P_\theta\}$ . Using the observation  $x$ , we make inferences about  $\theta$ . Often, many aspects of  $x$  provide no information about  $\theta$ . This means that  $x$  can be summarized.

**Example 2.1.** (Number of Heads)

Suppose we have  $n$  i.i.d. Bernoulli trials (for example  $n$  coin flips). We observe  $x = \langle x_1, x_2, \dots, x_n \rangle \in X = \{0, 1\}^n$ . Let the parameter  $\theta = p(1) = \pi$ .

Notice that the ordering of the outcomes is irrelevant: it gives no information about  $\pi$ . We can infer  $\pi$  just as well knowing only how many trials resulted in 1, i.e.  $S = \sum_{i=1}^n x_i$ . In the case of coin flips, we only care about the number of heads and not the result of the first toss, the second toss, etc.

**Definition 1.** A *statistic* is any function  $t : X \rightarrow Y$  for any range  $Y$ . Often  $Y = \mathbb{R}$ .

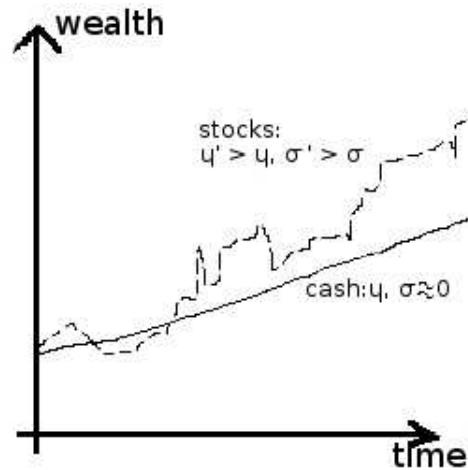


Figure 1: Stocks vs Cash

In the Bernoulli example, the counting function  $x \mapsto \sum_{i=1}^n x_i$  is a *statistic*. Note that estimators are also statistics. They are functions that depend only on the observed sample.

Every statistic is a summary of the observed sample. The statistic preserves some information and loses other information. Like an abstraction in software engineering, this loss of detail is sometimes useful. How is information lost? If  $t : X \rightarrow Y$  and  $t(x) = t(x')$  then we lose the distinction between  $x$  and  $x'$ . Since a statistic only depends on  $x$ , it is something that can actually be calculated from data, as opposed to a concept that we can only think about theoretically.

**Example 2.2.** Here are functions that we can only think about:

1. In general,  $g : \Theta \rightarrow \mathbb{R}$  is cannot be used in numerical calculations.
2. A more specific example:  $g : \langle \mu, \sigma^2 \rangle \mapsto \mu$
3. Another one, useful in finance:  $g' : \langle \mu, \sigma^2 \rangle \mapsto \frac{\mu}{\sigma}$

Figure 1 illustrates two investments with different  $\frac{\mu}{\sigma}$  ratios. A higher ratio is intuitively better since it translates to higher average returns with lower variance, where variance is a formalization of the intuitive concept of “risk.”

### 3 Sufficiency

Now we will introduce the concept of sufficiency. Intuitively, a statistic is *sufficient* if it preserves all information from  $x$  that is relevant for estimating which distribution  $P_\theta$  generated  $x$ . We will show that in the Bernoulli example,  $S = \sum_{i=1}^n x_i$  is sufficient for  $\pi$ .

First we need to divide the sample space  $X$  into partitions. Let  $\{A\}$  be a family of subsets that is a partition of  $X$ . For a given subset  $A$ :

$$\begin{aligned} P_\theta(x|x \in A) &= 0 \text{ if } x \notin A \\ &= \frac{P_\theta(x)}{P_\theta(A)} \text{ if } x \in A \end{aligned}$$

Assume that for every  $A$ ,  $P_{\theta_1}(x|x \in A) = P_{\theta_2}(x|x \in A)$  for all  $\theta_1$  and  $\theta_2$ . In general though,  $P_{\theta_1}(x) \neq P_{\theta_2}(x)$ .

**Example 3.1.** (Where the above assumption is true – Bernoulli)

Let  $x = \langle x_1, x_2, \dots, x_n \rangle \in X = \{0, 1\}^n$ .

Let the sample space be partitioned as  $X = \bigcup_{j=0}^n A_j$  where  $A_j$  is the event that we have  $j$  ones:  $A_j = \{x \in X \mid \sum_{i=1}^n x_i = j\}$ . Note that  $P_\theta(x|x \in A_k) = 0$  if  $\sum_{i=1}^n x_i \neq k$ . Also,

$$\begin{aligned} P_\theta(x) &= \prod_{i=1}^n [\pi \text{ if } x_i = 1 \text{ and } 1 - \pi \text{ if } x_i = 0] \\ &= \prod_{i=1}^n \pi^{x_i} (1 - \pi)^{(1-x_i)}. \end{aligned}$$

Therefore

$$\begin{aligned} P_\theta(x|x \in A_k) &= \prod_{i=1}^n \frac{\pi^{x_i} (1 - \pi)^{(1-x_i)}}{P(A_k)} \\ &= \frac{\pi^k (1 - \pi)^{(n-k)}}{\binom{n}{k} \pi^k (1 - \pi)^{(n-k)}} \\ &= \frac{1}{\binom{n}{k}} \text{ if } x \in A. \end{aligned}$$

Recall that our assumption was  $P_{\theta_1}(x|x \in A) = P_{\theta_2}(x|x \in A)$  for all  $A$  and all  $\theta_1$  and  $\theta_2$ . Notice that the above conditional probability does not depend on  $\theta = \pi$ . So, the Bernoulli distribution is an example where our assumption holds.

*Thoughts:* Sometimes we cannot observe  $x$  directly (fully), but we can observe that  $x \in A_k$ . Is that relevant for estimating  $\theta$ ? Yes. In general,  $P_{\theta_1}(A_k) \neq P_{\theta_2}(A_k)$ . On the other hand, what if we also discover *which*  $x \in A_k$  was the sample? This extra information is *not* helpful, if the earlier assumption is true.

**Definition 2.** (Sufficient partitions) The *partition*  $\{A\}$  of  $X$  is *sufficient* for the family of distributions  $P_\theta$  if  $P_\theta(x|x \in A)$  is the same for all  $\theta$  and for all  $A \in \{A\}$ .

**Definition 3.** The partition  $\{A\}$  is *minimal sufficient* if its sets are supersets of those of every other sufficient partition.

That is, a minimal sufficient partition is as coarse (the opposite of refined) as possible. For example,  $\{x\}$  is a partition but it is likely not minimal.

## 4 Achieving minimal sufficiency

We will define a particular equivalence relation:  $x \approx x'$  iff  $P_\theta(x)/P_\theta(x')$  is the same for all  $\theta$ . The following properties hold (by the definition of an equivalence relation):

1.  $\forall x \in X, x \approx x$  (reflexivity)
2.  $\forall x_1, x_2$ , if  $x_1 \approx x_2$  then  $x_2 \approx x_1$  (symmetry)
3.  $\forall x_1, x_2, x_3$ , if  $x_1 \approx x_2$  and  $x_2 \approx x_3$  then  $x_1 \approx x_3$  (transitivity)

**Example 4.1.** How does the above equivalence relation apply to the Bernoulli example?

$$\begin{aligned} \frac{P_\theta(x)}{P_\theta(x')} &= \frac{\pi^{\sum x_i} (1 - \pi)^{n - \sum x_i}}{\pi^{\sum x'_i} (1 - \pi)^{n - \sum x'_i}} \\ &= \pi^{\sum x_i - \sum x'_i} (1 - \pi)^{\sum x'_i - \sum x_i} \end{aligned}$$

which has no dependency on the parameter  $\pi$  iff  $\sum x_i = \sum x'_i$ .

The purpose of  $\approx$  is that equivalence relations define a *partition* of a set (where each subset contains equivalent elements).

**Lemma 4.1.** This equivalence relation  $\approx$  is minimal sufficient.

In the Bernoulli example, a partition based on  $\sum x_i$  is minimal sufficient.

**Definition 4.** The *statistic*  $t : X \rightarrow Y$  is (minimal) sufficient for  $P_\theta$  if the partition defined by the following equivalence relation is (minimal) sufficient:  $x \sim x'$  iff  $t(x) = t(x')$ .

## 5 Final notes on sufficiency

1. If  $t$  is a sufficient statistic and  $u$  is a minimal sufficient statistic, then  $u(x) = f(t(x))$  for some function  $f$ . Intuitively,  $f$  loses additional information to bring  $t(x)$  to the level of  $u(x)$ .
2. Minimal sufficient statistics are never unique. For example, if  $\sum x_i$  is minimal sufficient, then so is  $\bar{x} = \frac{\sum x_i}{n}$ . Both define the same partition of the sample space.
3. Sufficiency is relative to the family of distributions  $\{P_\theta\}$ . It is the same regardless of  $g(\theta)$ .