1 Principles of Data Reduction

In this chapter, we explore how we can use functions of a sample $\boldsymbol{X}$ to make inferences about an unknown parameter (of the distribution of the sample) θ.

Definition 1 (Statistic)**.** A statistic is any function of the data.

A statistic T forms a partition of the sample space $\mathcal{X}$ according to its image, $\mathcal{T} = \{t : \exists \boldsymbol{x} \in \mathcal{X} \text{ s.t. } t = T(\boldsymbol{x})\}$. In this way a statistic provides a method of data reduction. An experimenter who observes only T will treat as equal two samples $\boldsymbol{x}, \boldsymbol{y}$ for which $T(\boldsymbol{x}) = T(\boldsymbol{y})$.

1.1 The Sufficiency Principle

Definition 2 (Sufficient Statistic)**.** A statistic $T(\boldsymbol{X})$ is a *sufficient statistic* for θ if the conditional distribution of the sample $\boldsymbol{X}$ given $T(\boldsymbol{X})$ does not depend on θ.

Remark 1. We ignore the fact that all points have 0 probability for continuous distributions.

Definition 3 (The Sufficiency Principle)**.** If $T(\boldsymbol{X})$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\boldsymbol{X}$ only through $T(\boldsymbol{X})$.

Theorem 1. If $p(\boldsymbol{x}|\theta)$ is the pmf/pdf of the sample $\boldsymbol{X}$ and $q(t|\theta)$ is the pmf/pdf of $T(\boldsymbol{X})$, then $T(\boldsymbol{X})$ is a sufficient statistic for θ if $\forall \boldsymbol{x} \in \mathcal{X}$, $p(\boldsymbol{x}|\theta)/q(t|\theta)$ is constant as a function of θ.

Remark 2 (Niceness of the exponential family)**.** It turns out that outside of the exponential family, it is rare to have a sufficient statistic that is of smaller dimension than the size of the sample.

Theorem 2 (Factorisation Theorem)**.** Let $f(\boldsymbol{x}|\theta)$ denote the pmf/pdf of a sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is a sufficient statistic for θ if and only if there exists functions $g(t|\theta)$ and $h(\boldsymbol{x})$ such that $\forall \boldsymbol{x} \in \mathcal{X}$, $\forall \theta \in \Theta$

$$f(\boldsymbol{x}|\theta) = g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x}).$$

Remark 3. This theorem shows that the identity is a sufficient statistic. It is straightforward to show from this that any bijection of a sufficient statistic is a sufficient statistic.

Theorem 3. Let $X_1, \ldots, X_n$ be iid observations from a pmf/pdf $f(x|\theta)$ from an exponential family

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)\right),$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$, $d \leq k$. Then $\boldsymbol{T}(\boldsymbol{X})$ defined by

$$T_i(\boldsymbol{X}) = \sum_{j=1}^{k} t_i(\boldsymbol{X}_j)$$

is a sufficient statistic for $\boldsymbol{\theta}$.

Definition 4 (Minimal sufficient statistic)**.** A sufficient statistic $T(\boldsymbol{X})$ is called a *minimal sufficient statistic* if, for any other sufficient statistic $T'(\boldsymbol{X})$, T is a function of T'.

Remark 4. By 'function of' we mean that if $T'(\boldsymbol{x}) = T'(\boldsymbol{y})$ then $T(\boldsymbol{x}) = T(\boldsymbol{y})$ – T varies with respect to X only insofar as it varies with T'. This means that each tile of the partition of the sample space according to the image of T' is a subset of some tile in the partition according to T. This means that minimal sufficient statistics provide the coarsest possible tiling of the sample space and thus are the sufficient statistics that provide the greatest data reduction.

Theorem 4. Let $f(\boldsymbol{x}|\theta)$ be the pmf/pdf of a sample $\boldsymbol{X}$. Suppose there exists a function $T(\boldsymbol{X})$ such that $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ the ratio $f(\boldsymbol{x}|\theta)/f(\boldsymbol{y}|\theta)$ is constant as a function of θ if an only if $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Then $T(\boldsymbol{X})$ is a minimal sufficient statistic for θ.

Definition 5 (Necessary Statistic)**.** A statistic is *necessary* if it can be written as a function of every sufficient statistic.

Theorem 5. A statistic is minimal sufficient if and only if it is a necessary and sufficient statistic.

Definition 6 (Ancillary Statistic)**.** A statistic $S(\boldsymbol{X})$ whose distribution does not depend on the parameter θ is called an *ancillary statistic*.

Definition 7 (First Order Ancillary)**.** A statistic $V(\boldsymbol{X})$ is *first order ancillary* if $\mathbb{E}[V(\boldsymbol{X})]$ is independent of θ.

Definition 8 (Complete Statistic)**.** Let $f(t|\theta)$ be a family of pdfs or pmfs for a statistic $T(\boldsymbol{X})$. The family of probability distributions is called complete if

$$\mathbb{E}[g(T)] = 0 \ \forall \theta \implies \mathbb{P}(g(T) = 0) = 1 \ \forall \theta.$$

Equivalently, $T(\boldsymbol{X})$ is called a *complete statistic*.

(It is left unsaid that the function g must be independent of θ.)

Theorem 6 (Basu's Theorem)**.** If $T(\boldsymbol{X})$ is a complete and minimal sufficient statistic, then $T(\boldsymbol{X})$ is independent of every ancillary statistic.

Theorem 7 (Complete statistics in the exponential family)**.** Let $X_1, \ldots, X_n$ be iid observations from a pmf/pdf $f(x|\theta)$ from an exponential family

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right),$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$, $d \leq k$. Then $\boldsymbol{T}(\boldsymbol{X})$ defined by

$$T_i(\boldsymbol{X}) = \sum_{j=1}^{k} t_i(\boldsymbol{X}_j)$$

is complete if $\{(w_1(\boldsymbol{\theta}), \ldots, w_n(\boldsymbol{\theta}))\}$ contains an open set in $\mathbb{R}^k$.

Remark 5. The open set criteria excludes curved exponential families.

Theorem 8. If a minimal sufficient statistic exists, then every complete statistic is minimal sufficient.

1.2 The Likelihood Principle

Definition 9 (Likelihood)**.** Let $f(\boldsymbol{x}|\theta)$ denote the pmf/pdf of the sample $\boldsymbol{X}$. Then the *likelihood function*, given an observation $\boldsymbol{X} = \boldsymbol{x}$, is

$$L(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta)$$

as a function of θ.

Definition 10 (Likelihood Principle)**.** If $\boldsymbol{x}$ and $\boldsymbol{y}$ are two sample points such that $L(\theta|\boldsymbol{x})$ is proportional to $L(\theta|\boldsymbol{y})$, that is, there exists a constant $C(\boldsymbol{x}, \boldsymbol{y})$ such that

$$L(\theta|\boldsymbol{x}) = C(\boldsymbol{x}, \boldsymbol{y}) L(\theta|\boldsymbol{y}) \quad \forall \theta,$$

then the conclusions drawn from $\boldsymbol{x}$ and $\boldsymbol{y}$ should be identical.

1.3 A Slightly More Formal Construction

Definition 11 (Experiment)**.** We define an *experiment* E to be a triple $(\boldsymbol{X}, \theta, f(\boldsymbol{x}|\theta))$, where $\boldsymbol{X}$ is a random vector with pmf/pdf f.

Knowing what experiment E was performed, an experimenter will observer $\boldsymbol{X} = \boldsymbol{x}$. The conclusions they draw about θ will be denoted $\text{Ev}(E, \boldsymbol{x})$, which stands for *the evidence about* θ *arising from* E *and* $\boldsymbol{x}$.

Definition 12 (Formal Sufficiency Principle)**.** Consider an experiment $E = (\boldsymbol{X}, \theta, f(\boldsymbol{x}|\theta))$ and suppose that $T(\boldsymbol{X})$ is a sufficient statistic for θ. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are sample points satisfying $T(\boldsymbol{x}) = T(\boldsymbol{y})$, then $\text{Ev}(E, \boldsymbol{x}) = \text{Ev}(E, \boldsymbol{y})$.

Definition 13 (Conditionality Principle)**.** Suppose that $E_1 = \{X_1, \theta, f_1(x_1|\theta)\}$ and $E_2 = \{X_2, \theta, f_2(x_2|\theta)\}$ are two experiments, where only the unknown parameter θ need be common between the two experiments. Consider the mixed experiment in which the random variable J is observed, where $\mathbb{P}(J = 1) = \mathbb{P}(J = 2) = \frac{1}{2}$ (independent of $\boldsymbol{X}_1, \boldsymbol{X}_2, \theta$), and then the experiment E_J is performed. Formally, the experiment performed is $E^* = (\boldsymbol{X}^*, \theta, f(\boldsymbol{x}^*|\theta))$, where $\boldsymbol{X}^* = (j, \boldsymbol{X})j$ and $f^*(\boldsymbol{x}^*|\theta) = f^*((j, \boldsymbol{x}_j)|\theta) = \frac{1}{2} f_j(\boldsymbol{x}_j|\theta)$. Then

$$\text{Ev}(E^*, (j, \boldsymbol{x}_j)) = \text{Ev}(E_j, \boldsymbol{x_j}).$$

That is, information about θ depends only on the experiment run (not on the fact that the particular experiment was chosen).

Definition 14. Formal Likelihood Principle Suppose that we have two experiments, $E_1 = (\boldsymbol{X}_1, \theta, f_1(\boldsymbol{x}_1|\theta))$ and $E_2 = (\boldsymbol{X}_2, \theta, f_2(\boldsymbol{x}_2|\theta))$ where the unknown parameter θ is the same in both experiments. Suppose that $\boldsymbol{x}_1^*$ and $\boldsymbol{x}_2^*$ are sample points from E_1 and E_2 respectively, such that

$$L(\theta|\boldsymbol{x}_2^*) = CL(\theta|\boldsymbol{x}_1^*)$$

for all θ and for some constant $C(\boldsymbol{x}_1^*, \boldsymbol{x}_2^*)$ that is independent of θ. Then

$$\text{Ev}(E_1, \boldsymbol{x}_1^*) = \text{Ev}(E_2, \boldsymbol{x}_2^*).$$

Remark 6. Note that this is more general from the other likelihood principle since it concerns two experiments (that we can of course set to be equal).

Corollary 1 (Likelihood Principle Corollary)**.** If $E = (\boldsymbol{X}, \theta, f(\boldsymbol{x}|\theta))$ is an experiment then $\text{Ev}(E, \boldsymbol{x})$ should depend on E and $\boldsymbol{x}$ only through $L(\theta|\boldsymbol{x})$.

Theorem 9 (Birnbaum's Theorem)**.** The Formal Likelihood Principle follows from the Formal Sufficiency Principle and the Conditionality Principle. The converse is also true.

Remark 7. Many common statistical procedures violate the Formal Likelihood Principle – the topic of the applicability of these principles is not settled. For instance, checking the residuals of a model (to grade the model) violates the Sufficiency Principle. These notions are model dependent, so may not be applicable until *after* we have decided on a model.

1.4 The Equivariance Principle

Definition 15 (Measurement Equivariance)**.** Inferences should not depend on the measurement scale used.

Definition 16 (Formal Invariance)**.** If two inference problems have the same formal structure, in terms of the mathematical model used, then the same inference procedure should be used, regardless of the physical realisation.

Definition 17 (Equivariance Principle)**.** If $\boldsymbol{Y} = g(\boldsymbol{X})$ is a change of measurement scale such that the model $\boldsymbol{Y}$ has the same formal structure as the model $\boldsymbol{X}$, then an inference procedure should be both measurement equivariant and formally invariant.

Definition 18. Let $\mathcal{F} = \{f(\boldsymbol{x}|\theta) : \theta \in \Theta\}$ be a set of pdfs or pmfs for $\boldsymbol{X}$, and let $\mathcal{G}$ be group of transformations on the sample space $\mathcal{X}$. Then $\mathcal{F}$ is *invariant under the group* $\mathcal{G}$ if for every $\theta \in \Theta$ and $g \in \mathcal{G}$ there exists a unique $\theta' \in \Theta$ such that $\boldsymbol{Y} = g(\boldsymbol{X})$ has the distribution $f(\boldsymbol{y}|\theta')$ if $\boldsymbol{X}$ has the distribution $f(\boldsymbol{x}|\theta')$.