A Opening remarks and preliminaries

In previous probability theory coursework, we studied randomness using stochastic data-generating models. Given a random variable (observable) X with CDF, F, and PDF or PMF, f, we compute expressions such as an expectation $\mathbb{E}[X]$ or tail probability $\Pr[X > x]$.

Statistical inference consists of making probabilistic statements about some or all parts of a statistical model. In this course, we focus on making inferences about unknown quantities.

> *Models, of course, are never true, but fortunately it is only necessary that they be useful.*
>
> George Box (1979), founder of UW–Madison Statistics

A few comments regarding semantics:

- *Random* (as in *random variable*) is not equivalent in meaning to *arbitrary*.
- We specify (i.e., we model) *randomness* by using probability distributions and by making certain assumptions about *stochasticity*.
- *Stochastic* means "having an underlying probability distribution or pattern that may be analyzed statistically but may not be predicted precisely".

In this class, we often suppose that one or multiple random experiments are performed from which we collect observable data. Our objective is to extract information and meaning from data, to interpret our results, and to draw conclusions (i.e., to make inferences).

The observable data set is a realization of a random vector defined on an underlying sample space. We refer to the *population* as the distribution of the random vector, or in some cases the set of elements from which we draw a sample. The random vector that produces the data is called a *sample* from the population.

The size of the data set is called the *sample size*, typically denoted by n. We say that the population is known when the data-generating distribution is completely known. In most statistical settings, the population is only (at least) partially assumed known, and we wish to deduce remaining unknown properties of the population based on an available random sample. Informally, a statistical model is a set of assumptions on the population that often (i) makes analysis possible or somewhat easy and (ii) is based on problem-specific knowledge, experience, or expertise. Statistical models that are fully specified by a finite-dimensional vector of parameters are said to be *parametric*. Statistical models that are only partially specified by a parametric component are said to be *semiparametric*. Otherwise, a model is said to be *nonparametric*.

Important terminology and mathematical notation:

- When considering a random experiment, we let Ω denote our **sample space**.
- On this space, we define a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$.
- Given $\omega \in \Omega$, an outcome of the experiment, $\boldsymbol{X}(\omega)$ denotes our **data** or **observations**.
- Usually, we view $\boldsymbol{X}$ as a mapping from Ω to $\mathbb{R}^n$.
- We say that $\boldsymbol{X}$ is observable whereas $\boldsymbol{X} = \boldsymbol{x}$ is observed. Here, $\boldsymbol{x} = (x_1, \ldots, x_n)$.
- A family $\mathcal{P}$ of probability distributions (on $\mathbb{R}^n$) is called a **statistical model**.

A **parametrization** is a mapping $\theta \mapsto P_\theta$ from a parameter space (space of labels) $\Theta \ni \theta$ to a statistical model $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\}$. Conceptually, a parametrization assigns "labels" or "identifiers" to distributions in a statistical model.

A parametrization is **unidentifiable** if, for some θ_1, θ_2, we have $\theta_1 \neq \theta_2$ and yet $P_{\theta_1} = P_{\theta_2}$.

A parametrization is **identifiable** whenever $P_{\theta_1} = P_{\theta_2}$ implies $\theta_1 = \theta_2$. In other words, identifiable parametrizations are injective maps (i.e., one-to-one functions). Equivalently, $\theta_1 \neq \theta_2$ implies $P_{\theta_1} \neq P_{\theta_2}$.

Problem A.0.1. Which of the following parametrizations are identifiable? Justify your answers.

1. Let $X_1, \ldots, X_p$ be independent with $X_i \sim \mathcal{N}(\alpha_i + \nu, \sigma^2)$. Write

$$\boldsymbol{\theta} = (\alpha_1, \alpha_2, \ldots, \alpha_p, \nu, \sigma^2),$$

   and let $P_{\boldsymbol{\theta}}$ denote the distribution of $\boldsymbol{X} = (X_1, \ldots, X_p)$.

2. Same as (1) but with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ restricted to the set

$$\left\{ (a_1, \ldots, a_p) : \sum_{i=1}^{p} a_i = 0 \right\}.$$

3. Let X and Y be independent $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$, $\boldsymbol{\theta} = (\mu_1, \mu_2)$, and we observe $Y - X$.

4. Let X_{ij} for $i = 1, \ldots, p; j = 1, \ldots, b$ be independent with $X_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2)$, where $\mu_{ij} = \nu + \alpha_i + \lambda_j$, $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_p, \lambda_1, \ldots, \lambda_b, \nu, \sigma^2)$, and $P_{\boldsymbol{\theta}}$ is the distribution of $X_{11}, \ldots, X_{pb}$.

5. Same as (4) but with $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ satisfying the conditions $\sum_{i=1}^{p} \alpha_i = 0$ and $\sum_{j=1}^{b} \lambda_j = 0$.

A **parameter** ν ("nu") is a map $\nu : \mathcal{P} \to \mathcal{N}$, where $\mathcal{P}$ is a statistical model and $\mathcal{N}$ is another space. Conceptually, $\nu(P)$ is a feature of the distribution P generating the data X.

Lemma A.0.1. Given a parametrization $\theta \mapsto P_\theta$, θ is a parameter if and only if the parametrization is identifiable.

Proof. Exercise. □

Lemma A.0.2. The parameter $\theta : \mathcal{P} \to \Theta$ is well-defined as the inverse of the mapping $\theta \mapsto P_\theta$ from Θ to its range in $\mathcal{P}$ if and only if $P_{\theta_1} = P_{\theta_2}$ implies $\theta_1 = \theta_2$. Consequently, $\theta(P_\theta) = \theta$.

Proof. Exercise. □

Remark (Model representations). A parameter can have many different representations. For example, the mean, median, and center of symmetry coincide for Gaussian distributions. A vector-valued parametrization that is unidentifiable can still have components that are identifiable (hence, parameters).

Remark (Notation). Given a parameter value θ and distribution P_θ, we write $\mathbb{E}_\theta$ to denote the expectation under $X \sim P_\theta$. Cumulative distribution functions (CDFs) are written as $F(\cdot;\theta)$ or $F_\theta(\cdot)$ or $F_X(\cdot)$ or F, depending on context. Probability density or mass functions (PDFs and PMFs) are written as $f(\cdot;\theta)$ or $f_\theta(\cdot)$ or $f_X(\cdot)$ or f, depending on context. The notation p is sometimes used instead of f. The notation $\mid$ is sometimes used instead of a semicolon, namely $f(x \mid \theta)$ and $f(x;\theta)$, where both $\mid$ and ; should be read as "given", though the former is more frequently encountered when discussing conditional distributions.

Remark (Model properties). We say that $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\}$ is a **parametric model** if $\Theta \subseteq \mathbb{R}^d$ for some positive integer $d \geq 1$. We say that $\mathcal{P}$ is a **regular parametric model** if, in addition to being parametric, it consists of continuous distributions P_θ with densities $f(\boldsymbol{x};\theta)$ or discrete distributions with frequency functions $f(\boldsymbol{x};\theta)$ where the support $\{\boldsymbol{x} : f(\boldsymbol{x};\theta) > 0\}$ is the same set for all $\theta \in \Theta$.

B Properties of a random sample

Definition (Random sample). A collection of random variables (or vectors) $X_1, \ldots, X_n$ is called a **random sample** (of size n from a population determined by F) when (i) the random variables are independent and (ii) each observable X_i has cdf given by F, namely $X_i \sim F$. Equivalently, we say $X_1, \ldots, X_n$ form an IID sample.

We often write $\boldsymbol{X} = (X_1, \ldots, X_n)$. Given an IID sample $X_1, \ldots, X_n$, we often write X to denote a dummy random variable with the same distributional properties as each X_i.

Definition (Statistic). A **statistic** T is a function from the sample space $\mathcal{X}$ to some space of values $\mathcal{T}$. The probability distribution of a statistic T is called its **sampling distribution**.

Theorem 5.2.6. If $X_1, \ldots, X_n$ is a random sample from a population F on $\mathbb{R}$ with finite mean μ and variance σ^2, then, (a) $\mathbb{E}[\overline{X}_n] = \mu$, (b) $\mathrm{Var}[\overline{X}_n] = \sigma^2/n$, and (c) $\mathbb{E}[S_n^2] = \sigma^2$.

Proof. Exercise. □

A statistic T is said to be **unbiased** for a parameter θ if $\mathbb{E}[T] = \theta$. By the above theorem, $\overline{X} \equiv \overline{X}_n$ is unbiased for μ (for every choice of sample size $n \geq 1$) and $S^2 \equiv S_n^2$ is unbiased for σ^2 (for every choice of sample size $n \geq 2$). However, $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 = \frac{n-1}{n} S_n^2$ is not unbiased for σ^2 (for any choice of sample size $n \geq 2$).

A sequence of statistics T_n is said to be **asymptotically unbiased** for a parameter θ if $\lim_{n\to\infty} \mathbb{E}[T_n] = \theta$. Consequently, $\widehat{\sigma}_n^2$ above is asymptotically unbiased for σ^2.

Recall the following basic properties of (scaled) sample means.

Problem B.0.1. If $X_1, \ldots, X_n$ is a random sample from

$$\begin{bmatrix} \mathcal{N}(\mu, \sigma^2) \\ \text{Gamma}(\alpha, \beta) \\ \text{Poisson}(\lambda) \\ \text{Binomial}(m, p) \\ \text{Cauchy}(\mu, \sigma) \end{bmatrix}, \text{ then } \begin{bmatrix} \overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n) \\ \overline{X}_n \sim \text{Gamma}(n\alpha, \beta/n) \\ n\overline{X}_n \sim \text{Poisson}(n\lambda) \\ n\overline{X}_n \sim \text{Binomial}(nm, p) \\ \overline{X}_n \sim \text{Cauchy}(\mu, \sigma) \end{bmatrix}.$$

Can you derive similar results for other common distributions?

Many common distributions are also related to each other via transformations $Y = g(X)$. See, for example, C&B Exercise 3.24.

Definition (Location-scale families). A random sample $X_1, \ldots, X_n$ is from a population in a **location-scale family** when the PDF of X_i is of the form $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ with a known PDF f and parameters $\mu \in \mathbb{R}$, $\sigma > 0$.

Many commonly encountered probability distributions, despite their apparent differences (e.g., discrete versus continuous), can in fact be written and studied in a unified fashion.

Definition (Exponential families, multiparameter case). A random sample $X_1, \ldots, X_n$ is from a population in an **exponential family** when the PDF or PMF has the form

$$f(x;\theta) = h(x) \times c(\theta) \times \exp\left[\sum_{j=1}^{k} w_j(\theta) t_j(x)\right],$$

where h, c, w_j, and t_j are known functions, k is a fixed positive integer, and $\theta \in \Theta$ is a parameter vector in the parameter space $\Theta \subset \mathbb{R}^k$. Moreover, an exponential family is further said to have **full rank** if Θ contains an open set in $\mathbb{R}^k$.

Remark. For simplicity, consider one-parameter exponential families. In other references, it is more common to write

$$p(x;\theta) = h(x) \times \exp\left[\eta(\theta)T(x) - B(\theta)\right]$$

and to consider the "canonical" reparametrization in terms of $\eta \equiv \eta(\theta)$, thus giving

$$q(x;\eta) = h(x) \times \exp\left[\eta T(x) - A(\eta)\right], \qquad x \in \mathcal{X}. \tag{1}$$

The model (1) with $\eta \in \mathcal{E} = \{\eta : |A(\eta)| < \infty\}$ is called the **one-parameter canonical exponential family generated by T and h.** We call $\mathcal{E}$ the **natural parameter space** and call T the **natural sufficient statistic**.

Notice that if we have a random sample from an underlying one-parameter exponential family, then the joint density (frequency) continues to be an exponential family since

$$f(\boldsymbol{x};\theta) = \left[\prod_{i=1}^{n} h(x_i)\right] \exp\left[\eta(\theta) \sum_{i=1}^{n} T(x_i) - nB(\theta)\right].$$

Examples:

We will repeatedly revisit exponential families and their properties throughout this course.

Certain sampling distributions are easy to derive in the context of exponential families:

Theorem 5.2.11. Let $X_1, \ldots, X_n$ be a random sample from a PDF or PMF $f(x;\theta)$, where

$$f(x;\theta) = h(x) \times c(\theta) \times \exp\left(\sum_{j=1}^{k} w_j(\theta) t_j(x)\right)$$

is a member of an exponential family. Define the statistics $T_1, \ldots, T_k$ by

$$T_j(X_1, \ldots, X_n) = \sum_{i=1}^{n} t_j(X_i), \qquad j = 1, \ldots, k.$$

If the set $\{(w_1(\theta), \ldots, w_k(\theta)) : \theta \in \Theta\} \subset \mathbb{R}^k$ contains an open subset of $\mathbb{R}^k$, then the distribution of $(T_1, \ldots, T_k)$ is an exponential family of the form

$$f_T(u_1, \ldots, u_k;\theta) = H(u_1, \ldots, u_k) \times [c(\theta)]^n \times \exp\left(\sum_{j=1}^{k} w_j(\theta) u_i\right).$$

The following theorem is a conceptually and computationally useful result relating the function $A : \mathcal{E} \to \mathbb{R}$ in (1) to properties of the natural sufficient statistic T.

Theorem B.0.1. (Moment generating function in one-parameter exponential families). Suppose X is a random variable with density or mass function given by (1), and let $\eta \in \mathcal{E}$ be an interior point. Then, the moment generating function of $T(X)$ exists and is given by

$$M_{T(X)}(s) = \exp\left[A(s+\eta) - A(\eta)\right].$$

Moreover,

$$\mathbb{E}[T(X)] = A'(\eta), \qquad \text{Var}[T(X)] = A''(\eta).$$

Problem B.0.2. Let $X_1, \ldots, X_n$ be a random sample from a population with density

$$f(x;\theta) = \left(\frac{x}{\theta^2}\right) \exp\left[-\frac{x^2}{2\theta^2}\right], \qquad x > 0, \qquad \theta > 0.$$

(a) Find the natural sufficient statistic. (b) Compute the expectation and variance of the natural sufficient statistic, in terms of the specified parameter θ and separately in terms of the canonical parameter η.

Problem B.0.3. Show that $\mathcal{P}_\theta = \{\mathcal{N}(\mu, \sigma^2) : \theta = (\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$ is a two-parameter exponential family.

Every k-parameter exponential family is also a k'-dimensional exponential family with $k' > k$. Ponder this claim.

The minimal dimension of a exponential family is known as its **rank**. We say that an exponential family is of rank k if and only if the generating statistic (vector) $\boldsymbol{T}$ is k-dimensional and $1, T_1(X), \ldots, T_k(X)$ are linearly independent with positive probability, i.e., $\Pr_{\boldsymbol{\eta}}\left[a_0 + \sum_{j=1}^{k} a_j T_j(X) = 0\right] < 1$ unless all a_j are zero.

Problem B.0.4. Consider a random sample $X_1, \ldots, X_n$ from Multinomial$(1; \boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \mathcal{S}_{k-1}$ is a vector of strictly positive probabilities summing to one. Here, $\mathcal{S}_{k-1}$ denotes the unit simplex. Show that the multinomial family is rank $k-1$. (Use this opportunity to (re)familiarize yourself with the Multinomial distribution, a generalization of the Binomial distribution.)

Theorem B.0.2. (Characterization of rank k exponential families). Suppose that $\mathcal{P} = \{q(x, \boldsymbol{\eta}); \boldsymbol{\eta} \in \mathcal{E}\}$ is a canonical exponential family generated by $(\boldsymbol{T}_{k\times 1}, h)$ with open natural parameter space $\mathcal{E}$. The following statements are equivalent.

1. $\mathcal{P}$ is rank k;
2. $\boldsymbol{\eta}$ is a parameter (identifiable);
3. $\mathrm{Var}_{\boldsymbol{\eta}}[\boldsymbol{T}]$ is positive definite;
4. $\boldsymbol{\eta} \mapsto \dot{A}(\boldsymbol{\eta})$ is injective (1-1) on $\mathcal{E}$;
5. A is strictly convex on $\mathcal{E}$.

The proof of this theorem is beyond the scope of our class.

Corollary B.0.1. (More about exponential families) Assume the hypothesis of the previous theorem, and suppose $\mathcal{P}$ is rank k. Then,

- $\mathcal{P}$ may be parametrized by $\boldsymbol{\mu}(\boldsymbol{\eta}) \equiv \mathbb{E}_{\boldsymbol{\eta}}[\boldsymbol{T}]$ where $\boldsymbol{\mu}$ ranges over $\dot{A}(\boldsymbol{\mathcal{E}})$;
- $\log q(x, \boldsymbol{\eta})$ is a strictly concave function of $\boldsymbol{\eta}$ on $\mathcal{E}$.

See also Theorem 3.4.2 in Casella and Berger.

We briefly pause here to record several basic facts for Gaussian (normal) data.

Theorem 5.3.1* (Properties of Gaussian random samples). If $X_1, \ldots, X_n$ is a random sample from $\mathcal{N}(\mu, \sigma^2)$, then

1. $\overline{X}_n$ and S_n^2 are independent random variables,
2. $\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$,
3. $(n-1)S_n^2/\sigma^2$ has a central chi-squared distribution with $n-1$ degrees of freedom,
4. $\sqrt{n}(\overline{X}_n - \mu)/S_n$ has a central Student's t distribution with $n-1$ degrees of freedom,
5. $\sqrt{n}(\overline{X}_n - \mu)/S_n$ converges in distribution to a standard normal (Gaussian) random variable as $n \to \infty$,
6. Given another (independent) random sample $Y_1, \ldots, Y_m$ from $\mathcal{N}(\mu_Y, \sigma_Y^2)$, the expression $\frac{S_{X,n}^2/\sigma_X^2}{S_{Y,m}^2/\sigma_Y^2}$ has a central F distribution with degrees of freedom $n-1$ and $m-1$.

Proof. Exercise. □

Definition 5.4.1 (Order statistics). The **order statistics** of a random sample $X_1, \ldots, X_n$ are the sample values placed in ascending order. They are denoted by $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. More formally,

$$X_{(1)} = \min_{1 \le i \le n} X_i \qquad X_{(2)} = \text{second smallest } X_i \qquad \ldots \qquad X_{(n)} = \max_{1 \le i \le n} X_i.$$

Observe that both the sample mean and sample variance are functions of order statistics since $\sum_{i=1}^n X_i = \sum_{i=1}^n X_{(i)}$ and $\sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_{(i)}^2$.

The **sample range**, defined as $R = X_{(n)} - X_{(1)}$, is a measure of dispersion in the sample.

Define the "good (success) events" $\mathcal{E}_j = \{X_j \le x\}$. Let Y denote the number of successes in n trials. Recall that $Y \sim \text{Binomial}(n, F_X(x))$. We have

$$F_{X_{(j)}}(x) = \Pr[X_{(j)} \le x] = \Pr[Y \ge j] = \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k}.$$

By differentiation, the chain rule, and subsequent algebra, we obtain

$$f_{X_{(j)}}(x) = \frac{d}{dx} F_{X_{(j)}}(x) = \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} f_X(x) [F_X(x)]^{j-1} [1 - F_X(x)]^{n-j}.$$

Problem B.0.5. Recall that the density function for a exponential random variable with parameter $\lambda > 0$ is given by $f_X(x) = \lambda \exp(-\lambda x)$ for $x \ge 0$. Prove that if $X_1 \sim \text{Exponential}(\lambda_1)$ and $X_2 \sim \text{Exponential}(\lambda_2)$ are independent, then $X = \min\{X_1, X_2\}$ follows Exponential(λ) with $\lambda = \lambda_1 + \lambda_2$.

On your own and during TA office hours, be sure to review the following results and concepts.

Theorem 5.4.3. (Marginal) PMF for order statistics

Theorem 5.4.4. (Marginal) PDF for order statistics

Example 5.4.5. Uniform order statistic PDF

Theorem 5.4.6. (Joint) PDF for order statistics

Example 5.4.7. Distribution of the midrange and range statistic

(extra space)

Definition (Percentiles). For any fixed $p \in (0,1)$, the $(100p)$-th **sample percentile** is the observation such that (approximately) np of the observations are smaller than this observation and (approximately) $n(1-p)$ observations are larger than this observation:

$$\begin{cases} X_{(1)} & \text{if } p \leq (2n)^{-1} \\ X_{(\{np\})} & \text{if } (2n)^{-1} < p < 0.5 \\ X_{((n+1)/2)} & \text{if } p = 0.5 \text{ and } n \text{ is odd} \\ \left(X_{(n/2)} + X_{(n/2+1)}\right)/2 & \text{if } p = 0.5 \text{ and } n \text{ is even} \\ X_{(n+1-\{n(1-p)\})} & \text{if } 0.5 < p < 1 - (2n)^{-1} \\ X_{(n)} & \text{if } p \geq 1 - (2n)^{-1}. \end{cases}$$

Above, $\{b\}$ denotes the number b rounded to the nearest integer, namely $k-0.5 \leq b < k+0.5$ yields $\{b\} = k$.

- The **sample median** is the 50th sample percentile. It is a measure of location and possibly different from the sample mean.
- The **sample lower quartile** and **sample upper quartile** are the 25th and 75th sample percentile, respectively.
- The **sample mid-range** is defined as $V = \frac{X_{(1)}+X_{(n)}}{2}$.

We shall defer discussion of convergence concepts until our later treatment of asymptotics and large-sample approximations.

Recall from calculus that the **Taylor series expansion** of an infinitely differentiable function g about a point x_0 is of the form

$$\begin{aligned} g(x) &= g(x_0) + \frac{g^{(1)}(x_0)}{1!}(x - x_0) + \frac{g^{(2)}(x_0)}{2!}(x - x_0)^2 + \dots \\ &= \sum_{m=0}^{\infty} \frac{g^{(m)}(x_0)}{m!}(x - x_0)^m. \end{aligned}$$

Here, the m-th derivative is denoted by $g^{(m)}(x) = \frac{d^m}{dx^m} g(x)$.

Theorem 5.5.21 (Taylor). Suppose the r-th derivative of g evaluated at x_0 exists. Define $T_r(x) = \sum_{m=0}^{r} \frac{g^{(m)}(x_0)}{m!}(x - x_0)^m$. Then,

$$\lim_{x \to x_0} \frac{g(x) - T_r(x)}{(x - x_0)^r} = 0.$$

Informally, near x_0, $g(x) = T_r(x) + \text{Remainder}.$

More formally, $g(x) - T_r(x) = \int_{x_0}^{x} \frac{g^{(r+1)}(t)}{r!}(x - t)^r dt.$

In words, we view $T_r(x)$ as a (local) approximation of $g(x)$.

C Principles of data reduction

Given a sample $X_1, \ldots, X_n$ from a population, we wish to make inferences about a parameter θ. Informally, a sufficient statistic for a parameter θ is a statistic that captures all the information about θ contained in the sample. Any additional information in the sample, besides the value of the sufficient statistic, does not contain any more information about θ.

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 the value $T(\boldsymbol{X})$. In other words, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are two sample points such that $T(\boldsymbol{x}) = T(\boldsymbol{y})$, then the inference about θ should be the same whether $\boldsymbol{X} = \boldsymbol{x}$ or $\boldsymbol{X} = \boldsymbol{y}$ is observed.

Let $\mathcal{T} = \{t : t = T(\boldsymbol{x}) \text{ for some } \boldsymbol{x} \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under $T(\boldsymbol{x})$. Then, $T(\boldsymbol{x})$ partitions the sample space into sets $A_t = \{\boldsymbol{x} : T(\boldsymbol{x}) = t\}$ for $t \in \mathcal{T}$. The statistic summarizes the data in that, rather than reporting the entire sample $\boldsymbol{x}$, it reports only $T(\boldsymbol{x}) = t$, equivalently $\boldsymbol{x} \in A_t$.

Definition 6.2.1. A statistic $T(\boldsymbol{X})$ is a **sufficient statistic** for θ if the conditional distribution of the sample $\boldsymbol{X}$ given the value of $T(\boldsymbol{X})$ does not depend on θ.

Procedurally, it is much easier to establish sufficiency of given statistics using the following two theorems.

Theorem 6.2.2 (Determining sufficient statistics). If $p(\boldsymbol{x};\theta)$ is the joint PDF or PMF of $\boldsymbol{X}$ and $q(t;\theta)$ is the PDF or PMF of $T(\boldsymbol{X})$, then $T(\boldsymbol{X})$ is a sufficient statistic for θ if, for every $\boldsymbol{x}$ in the sample space, the ratio $p(\boldsymbol{x};\theta)/q(T(\boldsymbol{x});\theta)$ is a constant function of θ.

Proof. Self-study. □

Examples:

Theorem 6.2.6 (Factorization theorem). Let $f(\boldsymbol{x};\theta)$ denote the joint PDF or PMF for a sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is a sufficient statistic for θ if and only if there exist functions $h(\boldsymbol{x})$ and $g(t;\theta)$ such that, for all sample points $\boldsymbol{x}$ and all parameter values θ, the joint density admits a factorization of the form

$$f(\boldsymbol{x};\theta) = h(\boldsymbol{x}) \times g(T(\boldsymbol{x});\theta). \tag{2}$$

Proof. Self-study. □

Example 6.2.8 (Uniform sufficient statistic). Let $X_1, \ldots, X_n$ be IID Uniform$\{1, \ldots, \theta\}$ with PMF $f(x;\theta) = \theta^{-1} \cdot I(x \in \{1, \ldots, \theta\})$. Let $\mathcal{N}$ denote the set of natural numbers, and let $\mathcal{N}_\theta$ denote the first θ natural numbers. Define $T(\boldsymbol{x}) = \max_i x_i$. Then,

$$f(\boldsymbol{x};\theta) = \prod_{i=1}^{n} \theta^{-1} I_{\mathcal{N}_\theta}(x_i) = \theta^{-n} \prod_{i=1}^{n} I_{\mathcal{N}_\theta}(x_i) = \theta^{-n} \times \left(\prod_{i=1}^{n} I_{\mathcal{N}}(x_i) \right) \times I_{\mathcal{N}_\theta}(T(\boldsymbol{x})).$$

Example 6.2.9* (Normal sufficient statistic, both parameters unknown). Consider a random sample $X_1, \ldots, X_n$ from $\mathcal{N}(\mu, \sigma^2)$. Verify the following claims.

- If $\boldsymbol{\theta} = (\mu, \sigma^2)$ is unknown, then $T(\boldsymbol{X}) = (\overline{X}_n, S_n^2)$ is a two-dimensional sufficient statistic for $\boldsymbol{\theta}$.
- If σ^2 is known, then $\overline{X}_n$ is a sufficient statistic for μ.
- If μ is known, then $\sum_{i=1}^n (X_i - \mu)^2$ is a sufficient statistic for σ^2 (ponder scaling).
- Show that the preceding statement is not true if we replace μ by $\overline{X}_n$ in the statistic.

Let θ be a parameter (vector), and let η be a subset of components of θ. If T is sufficient for θ, then it is also sufficient for η (i.e., sufficiency of sub-families).

It follows from the Factorization Theorem that, if T is sufficient and U is a one-to-one function of T, then U is also sufficient (for what?)

Exponential families of distributions readily reveal sufficient statistics for their parameters.

Problem C.0.1. State and prove C&B Theorem 6.2.10. Does this result apply to the previous two examples?

Multiple sufficient statistics may exist for the same parameter in a specific problem. Which sufficient statistic should we prefer?

Definition (Minimal sufficient statistic). A sufficient statistic $T(\boldsymbol{X})$ is called a **minimal sufficient statistic** if, for any other sufficient statistic $T'(\boldsymbol{X})$, it holds that $T(\boldsymbol{x})$ is a function of $T'(\boldsymbol{x})$, i.e., if there exists $g(\cdot)$ such that $T(\boldsymbol{x}) = g(T'(\boldsymbol{x}))$.

In words, a minimal sufficient statistic achieves the greatest possible data reduction for a sufficient statistic. It yields the coarsest possible underlying partition for a sufficient statistic.

Theorem 6.2.13. Let $f(\boldsymbol{x};\theta)$ be the PDF or PMF of a sample $\boldsymbol{X}$. Suppose there exists a function $T(\boldsymbol{x})$ such that, for every two sample points $\boldsymbol{x}$ and $\boldsymbol{y}$, the ratio $f(\boldsymbol{x};\theta)/f(\boldsymbol{y};\theta)$ is constant as a function of θ if and only if $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Then $T(\boldsymbol{X})$ is a minimal sufficient statistic for θ.

Proof. Self-study. □

Example 6.2.15 (Uniform minimal sufficient statistic). Consider a random sample $X_1, \ldots, X_n \sim \text{Uniform}(\theta, \theta + 1)$ where $\theta \in \mathbb{R}$. We have

$$f(\boldsymbol{x}; \theta) = \prod_{i=1}^{n} I(\theta < x_i < \theta + 1) = I\left(\max_i x_i - 1 < \theta < \min_i x_i\right).$$

We claim that $f(\boldsymbol{x}; \theta)/f(\boldsymbol{y}; \theta)$ will be positive and constant as a function of θ if and only if $\min_i x_i = \min_i y_i$ and $\max_i x_i = \max_i y_i$. By the preceding theorem, $T(\boldsymbol{X}) = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic for θ.

Problem C.0.2. Consider a random sample $X_1, \ldots, X_n$ from $\mathcal{N}(\mu, \sigma^2)$.

1. Suppose σ^2 is known. First, show that $(\overline{X}_n, S_n^2)$ is a sufficient statistic (vector) for μ. Second, determine a minimal sufficient statistic for μ.
2. Suppose both μ and σ^2 are unknown. Determine a minimal sufficient statistic for (μ, σ^2). Can you obtain a second minimal sufficient statistic?

Problem C.0.3. Prove that any one-to-one function of a minimal sufficient statistic is also a minimal sufficient statistic. Use this result to show that $\widetilde{T}(\boldsymbol{X}) = (X_{(n)} - X_{(1)}, (X_{(n)} + X_{(1)})/2)$ is also minimal sufficient in the preceding example.

As we have seen, sufficient statistics capture all the information about a parameter that is contained in a sample. Conversely, we might be interested in statistics that contain no information about a parameter of interest.

Definition 6.2.16 (Ancillary statistics). A statistic $S(\boldsymbol{X})$ whose distribution does not depend on the parameter θ is called an **ancillary statistic**.

Problem C.0.4. Consider a random sample $X_1, X_2 \sim \mathcal{N}(\mu, 1^2)$. Determine one or more ancillary statistics for μ.

Example 6.2.17 (Uniform ancillary statistic). Consider a random sample $X_1, \ldots, X_n \sim \text{Uniform}(\theta, \theta + 1)$ where $\theta \in \mathbb{R}$. Define the range statistic $R \equiv R(\boldsymbol{X}) = X_{(n)} - X_{(1)}$. Show that $R \sim \text{Beta}(\alpha = n - 1, \beta = 2)$. The distribution of R does not depend on θ, hence by definition R is an ancillary statistic for θ.

Example 6.2.18 (Location family ancillary statistic). Show that the range statistic $R = X_{(n)} - X_{(1)}$ is an ancillary statistic for θ whenever $X_1, \ldots, X_n$ is a random sample from a location parameter family with CDF $F(x - \theta), \theta \in \mathbb{R}$.

How are ancillary statistics and (minimal) sufficient statistics related? Are ancillary statistics {always, sometimes, never} independent of (minimal) sufficient statistics?

Definition 6.2.21 (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 $\mathbb{E}_\theta[g(T)] = 0$ for all θ implies $\Pr_\theta[g(T) = 0] = 1$ for all θ. Here, $T(\boldsymbol{X})$ is called a **complete statistic**.

Verifying that a statistic is complete can require additional mathematical or technical observations, as evidenced in the following examples.

The following theorem shows that completeness, as an additional condition, is enough to guarantee that a minimal sufficient statistic is independent of any ancillary statistic.

Theorem 6.2.24 (Basu's theorem). If $T(\boldsymbol{X})$ is a complete and minimal sufficient statistic, then $T(\boldsymbol{X})$ is independent of every ancillary statistic.

The proof of Basu's theorem does not directly use the property of being "minimal" sufficient. It turns out that complete statistics are, in fact, minimal, so we must consider existence.

Theorem 6.2.28 (Cousin to Basu's theorem). If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.

Complete statistics are straightforward to identify in exponential families. Below, the open set condition excludes so-called *curved exponential families* such as $\mathcal{N}(\theta, \theta^2)$.

Theorem 6.2.25 (Complete statistics in exponential families). Let $X_1, \ldots, X_n$ be a random sample from an exponential family with PDF or PMF given by

$$f(x; \boldsymbol{\theta}) = h(x) \times c(\boldsymbol{\theta}) \times \exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta}) t_j(x)\right],$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Then, the statistic

$$T(\boldsymbol{X}) = \left(\sum_{i=1}^{n} t_1(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)$$

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

Proof. Self-study. □