Chapter 1

Statistical Modeling

1.1 Statistical Models

Example 1: (Sampling inspection). A lot contains $N$ products with defective rate $\theta$. Take a sample without replacement of $n$ products and get $x$ defective products. What are the defective rates?

Possible outcomes: GGDGGGGDD \ldots, realization of outcomes.

How do we connect the sample with the population?

Modelling — think of data as a realization of a the random experiment.
Observe that a ”D” $\Rightarrow$ $\theta$ is large,

a ”G” $\Rightarrow$ $\theta$ is small.

**Probability Law**: Under this physical experiment

$$P(X = x) = \frac{\binom{N\theta}{x} \binom{N - N\theta}{n - x}}{\binom{N}{n}},$$

for $\max(0, n - N(1 - \theta)) \leq x \leq \min(n, N\theta)$. Convention: $\binom{n}{0} = 1$, $\binom{n}{m} = 0$ if $m > n$.

For example, $X/n \approx \theta$ and

$$\sqrt{n}(X/n - \theta) \to N(0, \theta(1 - \theta)).$$
Parameter: \( \theta \) — unknown, fixed.

**Parameter space** \( \Theta \): the possible value of \( \theta \): \( \Theta = \{0/N, 1/N, \cdots, N/N \} \) or \([0, 1]\).

For this specific example, the model comes from physical experiment. Now suppose that \( N = 10,000 \), \( n = 100 \) and \( x = 2 \). Our problem becomes an inverse problem: What is the value of \( \theta \)?

Logically, if \( \theta = 1\% \), it is possible to get \( x = 2 \). If \( \theta = 2\% \), it is also possible to get \( x = 2 \). If \( \theta = 3.5\% \), it is also possible to get \( x = 2 \). So, given \( x = 2 \), we can not tell exactly which \( \theta \) it is. Our conclusion can not be drawn without uncertainty. However, we do know some are more likely than the others and the degree of uncertainty gets smaller, as \( n \) gets large, whatever \( N \) is.

**Summary:**

— Statisticians think data as realizations from a stochastic model; this connects
the sample and parameters.

— Statistical conclusions can not be drawn without uncertainty, as we have only a finite sample.

— Probability is from a box to sample, while statistics is from a sample to a box.

**Example 2**: A measurement model (e.g. molecular weight, RNA/protein expression level, fat-free weight). An object is weighed $n$ times, with outcomes $x_1, \cdots, x_n$. Let $\mu$ be the true weight. We think the observed data as realizations of random variables $X_1, \cdots, X_n$, modeled as

$$X_i = \mu + \varepsilon_i$$

where $\varepsilon_i$ is error of measurement noise.

**Assumptions**

i) $\varepsilon_i$ is independent of $\mu$.

ii) $\varepsilon_i, i = 1, 2, \cdots, n$ are independent.
iii) \( \varepsilon_i, i = 1, 2, \cdots, n \) are identically distributed.

iv) the distribution of \( \varepsilon \) is continuous, with \( E(\varepsilon) = 0 \); or specifically symmetric about 0: \( f(y) = f(-y) \) for any \( y \).

Often, we assume further that \( \varepsilon_i \sim N(0, \sigma^2) \). Parameters in the model \( \theta = (\mu, \sigma^2) \), where \( \sigma^2 \) is a nuisance parameter.

Given a realization \( \mathbf{x} = (x_1, \cdots, x_n) \) of \( \mathbf{X} = (X_1, \cdots, X_n) \), what is the value of \( \mu \)?
Logically, if $\mu = 100$, it is possible to observe $x$. If $\mu = 1$, it is also possible to observe $x$. So we can not absolutely tell what value of $\mu$ is. But from the square-root law:

$$\text{var}(\bar{X}) = E((\bar{X} - \mu)^2) = \frac{\sigma^2}{n}. $$

Thus, $\bar{x}$ is likely close to $\mu$ when $n$ is large.

![Figure 1.3: Distributions of individual observation versus that of average](image)

**Example 3**: Drug evaluation (Hypertension drug)

Drug A $\rightarrow$ m patients  \quad Drug B $\rightarrow$ n patients
Measurement: blood pressure.

To eliminate confounding factors, use randomized controlled experiment. Here are the hypothetical outcomes:

<table>
<thead>
<tr>
<th>Drug A</th>
<th>Drug B</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>120</td>
</tr>
<tr>
<td>110</td>
<td>140</td>
</tr>
<tr>
<td>160</td>
<td>160</td>
</tr>
<tr>
<td>187</td>
<td>180</td>
</tr>
<tr>
<td>153</td>
<td>133</td>
</tr>
<tr>
<td>x_1</td>
<td>y_1</td>
</tr>
<tr>
<td>x_2</td>
<td>y_2</td>
</tr>
<tr>
<td>x_3</td>
<td>y_3</td>
</tr>
<tr>
<td>x_4</td>
<td>y_4</td>
</tr>
<tr>
<td>x_5</td>
<td>y_5</td>
</tr>
<tr>
<td>y_6</td>
<td>y_6</td>
</tr>
</tbody>
</table>

To model the outcomes, a possible idealization is the following box-model.

Figure 1.4: Illustration of a two-sample problem

<table>
<thead>
<tr>
<th>Drug A</th>
<th>Drug B</th>
</tr>
</thead>
<tbody>
<tr>
<td>random outcomes</td>
<td>( X_1, \ldots, X_m )</td>
</tr>
<tr>
<td>realizations</td>
<td>( x_1, \ldots, x_m )</td>
</tr>
</tbody>
</table>
Further, we might assume that

\[ X_1, \cdots, X_m \overset{i.i.d.}{\sim} N(\mu_A, \sigma^2_A) \quad Y_1, \cdots, Y_n \overset{i.i.d.}{\sim} N(\mu_B, \sigma^2_B). \]

We sometimes assume further \( \sigma_A = \sigma_B = \sigma. \)

Parameters in the model: \( \theta = (\mu_A, \mu_B, \sigma_A, \sigma_B). \)

Parameters of interest: \( \mu = \mu_A - \mu_B \)
and possibly \( \sigma. \)

Connection sample with population: data are realizations from a population, whose distribution depends on \( \theta. \)

**Model diagnostics:** Statistical models are idealizations, postulated by statisticians — needed to be verified. For example, the data histograms should look like theoretical distributions. Two sample variances are about the same, etc.

**General formulation**

**Data:** \( \mathbf{x} = (x_1, \cdots, x_n) \) are thought of the realization of a random vector \( \mathbf{X} = (X_1, \cdots, X_n). \)
Model: The distribution of $X$ is assumed in $\mathcal{P} = \{ P_\theta : \theta \in \Theta \}$, $\Theta$ is the parametric space.

Objectives: Inferences about $\theta$.

— In Example 1:

$$P_\theta(x) = \frac{(N\theta)^x (N-N\theta)^{n-x}}{n^n},$$

where $\Theta = \{0, 1/N, \cdots, N/N\}$ or $[0, 1]$.

— In Example 2:

$$P_\theta(x) = \prod_{i=1}^{n} \sigma^{-1} \varphi \left( \frac{x_i - \mu}{\sigma} \right)$$

where $\varphi(\cdot)$ is the normal density, $\Theta = \{(\mu, \sigma), \mu > 0, \sigma > 0\}$.

— In Example 3:

$$P_\theta(x) = \prod_{i=1}^{m} \sigma_A^{-1} \varphi \left( \frac{x_i - \mu_A}{\sigma_A} \right) \prod_{i=1}^{n} \sigma_B^{-1} \varphi \left( \frac{y_i - \mu_B}{\sigma_B} \right),$$

where $\varphi(\cdot)$ is the normal density, $\Theta = \{(\mu_A, \mu_B, \sigma_A, \sigma_B) : \mu_A, \mu_B, \sigma_A, \sigma_B > 0\}$. 
— Data $\mathbf{x}$ or its random variable $\mathbf{X}$ can include both $x$- and $y$-component.

The parameter $\theta$ doesn’t have to be in $\mathbb{R}^k$. In Example 2, without the normality assumption,

$$P_{\theta}(\mathbf{x}) = \prod_{i=1}^{n} f(x_i - \mu),$$

assuming that $\{\varepsilon_i, i = 1, \cdots, n\}$ are i.i.d random variables with density $f$. Then,

$$\Theta = \{(\mu, f) : \mu > 0, f \text{ is symmetric}\}.$$  

Since no form of $f$ has been imposed, i.e. $f$ has not been parameterized, the parameter space $\Theta$ is called nonparametric or semiparametric.

**Basic assumption:** Throughout this class, we will assume that

(i) Continuous variables: All $P_{\theta}$ are continuous with densities $p(\mathbf{x}, \theta)$ or

(ii) Discrete variable: All $P_{\theta}$ are discrete with frequency functions $p(x, \theta)$. Further,

there exists a set $\{\mathbf{x}_1, \mathbf{x}_2, \cdots\}$ such that

$$\sum_{i=1}^{\infty} p(\mathbf{x}_i, \theta) = 1,$$

where $x_i$ is independent of $\theta$.  

For convenience, we will call \( p(x, \theta) \) as density in both cases.

**Identifiability of parameters:** There are sometimes more than one way of parameterization. In Example 3: write

\[
X_1, \ldots, X_m \overset{i.i.d}{\sim} N(\mu + \alpha_1, \sigma^2) \quad Y_1, \ldots, Y_n \overset{i.i.d}{\sim} N(\mu + \alpha_2, \sigma^2).
\]

\( \theta = (\mu, \alpha_1, \alpha_2, \sigma) \). Hence,

\[
p_{\theta}(x, y, \theta) = \prod_{i=1}^{m} \sigma^{-1} \varphi \left( \frac{x_i - \mu - \alpha_1}{\sigma} \right) \prod_{i=1}^{n} \sigma^{-1} \varphi \left( \frac{y_i - \mu - \alpha_2}{\sigma} \right),
\]

If \( \theta_1 = (0, 1, 2, 1) \) and \( \theta_2 = (0.5, 0.5, 1.5, 1) \), then \( P_{\theta_1} = P_{\theta_2} \). Thus, the parameters \( \theta \) are not identifiable.

**Identifiability:** The model \( \{ P_{\theta}, \theta \in \Theta \} \) is identifiable if \( \theta_1 \neq \theta_2 \) implies \( P_{\theta_1} \neq P_{\theta_2} \).

**Example 4:** (Regression Problem). Suppose a sample of data \( \{(x_{i1}, \ldots, x_{ip}, y_i)\}_{i=1}^{n} \) are collected e.g.

\[
y = \text{salary, } x_1 = \text{age, } x_2 = \text{year of experience, } x_3 = \text{job grade, } x_4 = \text{gender, } x_5 = \text{PC job}.
\]
We wish to study the association between $Y$ and $X_1, \cdots, X_p$. How to predict $Y$ based on $\mathbf{X}$? Any gender discrimination? (Note: the data $\mathbf{x}$ in the general formulation now include all $\{(x_{i1}, \cdots, x_{ip}, y_i)\}_{i=1}^n$).

— Model I: linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_5 X_5 + \varepsilon, \quad \varepsilon \sim G,$$

where $\varepsilon$ is the part that can not be explained by $\mathbf{X}$. Thus the parameter space is $\Theta = \{(\beta_0, \beta_1, \cdots, \beta_5, G)\}$.

— Model II: semiparametric model

$$Y = \mu(X_1, X_2, X_3) + \beta_4 X_4 + \beta_5 X_5 + \varepsilon.$$

The parameter space is $\Theta = \{ (\mu(\cdot), \beta_4, \beta_5, G) \}$.

— Model III: nonparametric model

$$Y = \mu(X_1, \cdots, X_5) + \varepsilon.$$
The parameter space is $\Theta = \{(\mu(\cdot), G)\}$.

**Modeling:** Data are thought of a realization from $(Y, X_1, \cdots, X_5)$ with the relationship between $X$ and $Y$ described above.

From this example, the model is a convenient assumption made by data analysts. Indeed, statistical models are frequently useful fictions. There are trade-offs among the choice of statistical models:

larger model $\Rightarrow$ reducing model biases

$\Rightarrow$ increasing estimation variance.

The decision depends also available sample size $n$.

**Statistics:** a function of data only, e.g.

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}, \quad X_1, \quad X_1^2 + \sqrt{X_2^2 + X_3^2 + 3},$$

but

$$X_1 + \sigma, \quad \overline{X} + \mu$$

are not.
**Estimator**: an estimating procedure for certain parameters, e.g. $\bar{X}$ for $\mu$.

**Estimate**: numerical value of an estimator when data are observed, e.g. 

$$n = 3, \bar{x} = \frac{2 + 6 + 4}{3} = 3.$$ 

Estimator — for all potential realizations, estimate — for a realized result.

**Note**: An estimator is an estimating procedure. The performance criteria for a method is based on estimator, while statistical decisions are based on estimate in real applications.

1.2 Bayesian Models

**Probability**: Two view points:

\[
\begin{align*}
\text{long run relative frequency} & \quad \text{Frequentist} \\
\text{prior knowledge w/brief} & \quad \text{Bayesian}
\end{align*}
\]
So far, we have assumed no information about $\theta$ beyond that provided by data. Often, we can have some (vague) knowledge about $\theta$. For example,

— defective rate is $1\%$

— the distribution of DNA nucleotides is uniform,

— the intensity of an image is locally corrected.

**Example 1.** (Continued) Based on past records, one can construct a distribution of defective rate $\pi(\theta)$:

$$P(\theta = i/N) = \pi_i, \quad i = 1, 2, \ldots, N.$$ 

This provides as a prior distribution. The defective rate $\theta_0$ of the current lot is thought of as a realization from $\pi(\theta)$. Given $\theta_0$,

$$P(X = x|\theta_0) = \binom{N\theta_0}{x}\binom{N-N\theta_0}{n-x}\binom{N}{n},$$

**Basic element of Bayesian models**
(i) The knowledge about $\theta$ is summarized by $\pi(\theta)$ — prior dist.

(ii) A realization $\theta$ from $\pi(\theta)$ serves as the parameter of $X$.

(iii) Given $\theta$, the observed data $x$ are a realization of $p_\theta$. The joint density of $(\theta, X)$ is $\pi(\theta)p(x|\theta)$.

(iv) The goal of the Bayesian analysis is to modify the prior of $\theta$ after observing $x$:

$$
\pi(\theta|X = x) = \begin{cases} 
\frac{\pi(\theta)p(x|\theta)}{\int \pi(\theta)p(x|\theta) \, d\theta}, & \theta \text{ continuous}, \\
\frac{\pi(\theta)p(x|\theta)}{\sum_\theta \pi(\theta)p(x|\theta)}, & \theta \text{ discrete}
\end{cases}
$$

  e.g. summarizing the distribution by posterior mean, median and SD, etc.
Example 5 (Quality inspection) Suppose that from the past experience, the defective rate is about 10%. Suppose that a lot consists of 100 products, whose quality is independent of each other.
The prior distribution about the lot’s defective rate is

\[ \pi(\theta_i) = P(\theta = \theta_i) = \binom{100}{i} 0.1^i 0.9^{100-i}, \quad \theta_i = \frac{i}{100}. \]

Prior mean and variance are

\[ E\theta = E\frac{X}{100} = 0.1 \]
\[ \text{var}(\theta) = \frac{1}{100^2} \text{var}(X) = \frac{100 \times 0.9 \times 0.1}{100^2}, \]
\[ SD(\theta) = 0.03. \]
Now suppose that \( n = 19 \) products are sampled and \( x = 10 \) are defective. Then

\[
\pi(\theta_i|X = 10) = \frac{P(\theta = \theta_i, X = 10)}{P(X = 10)} = \frac{\pi(\theta_i)P(X = 10|\theta = \theta_i)}{\sum_j \pi(\theta_j)P(X = 10|\theta = \theta_j)}.
\]

e.g.

\[
P(\theta \geq 0.2|X = 10) = P(100\theta - X \geq 10|X = 10)
\approx 1 - \Phi\left(\frac{10 - 81 \times 0.1}{\sqrt{81 \times 0.9 \times 0.1}}\right)
\approx 30%.
\]

(100\theta - X is the number of defective left after 19 draws, having distribution Bernoulli(81, 0.1)). Compared with the prior probability

\[
P(\theta \geq 0.2) = P(100\theta \geq 20)
= 1 - \Phi\left(\frac{20 - 100 \times 0.1}{\sqrt{100 \times 0.9 \times 0.1}}\right)
\approx 0.1%,
\]

where 100\theta \sim Bernoulli(100,0.1).

**Example 6.** Suppose that \( X_1, \cdots, X_n \) are i.i.d. random variables with Bernoulli(\( \theta \)) and \( \theta \) has a prior distribution \( \pi(\theta) \). Then

\[
\pi(\theta|x) = \frac{\pi(\theta)\theta^{\sum_{i=1}^{n}x_i}(1 - \theta)^{n - \sum_{i=1}^{n}x_i}}{\int_0^1 \pi(t)t^{\sum_{i=1}^{n}x_i}(1 - t)^{n - \sum_{i=1}^{n}x_i} dt}.
\]
Figure 1.8: Beta distributions with shape parameters: Left panel: (4, 10), (5, 2), (2, 5), (0.7, 3); right panel: (5, 5), (2, 2), (1, 1), (0.5, 0.5)

If $\theta \sim \text{Beta}(r, s)$, i.e.

$$
\pi(\theta) = \frac{\theta^{s-1}(1-\theta)^{r-1}}{B(s, t)}, \quad E\theta = \frac{s}{r + s},
$$
then
\[ \pi(\theta|x) \propto \theta^{s + \sum x_i - 1}(1 - \theta)^{n - \sum x_i + r} \sim \text{Beta}(s + \sum x_i, n - \sum x_i + r). \]

Thus,
\[
E(\theta|x) = \frac{s + \sum_{i=1}^{n} x_i}{n + s + r} = \begin{cases} 
\frac{\sum_{i=1}^{n} x_i + 1}{n + 2} & s = r = 1 \\
\approx n^{-1} \sum_{i=1}^{n} x_i , & n \text{ is large}
\end{cases}
\]

**Conjugate prior**: Note that the prior and posterior in this example belong to the same family. Such a prior is called “conjugate prior”. It was introduced to facilitate the computation.

### 1.3 Sufficiency

Commonly-used principles for data reduction

\[
\begin{align*}
1^o & \quad \text{Sufficiency} \\
2^o & \quad \text{Invariant/equivariant}
\end{align*}
\]
Purpose:

1. simplify probability structure, less obscure than the whole data
2. understand whether a loss in reduction
3. useful technical tools

Example 7. A machine produces \( n \) items in secession with probability \( \theta \) of producing defective product. Suppose that there is no dependence between the quality of products.

![Probability model and its summary statistic.](image)

Figure 1.9: Probability model and its summary statistic.
Then, the probability model is

$$p(x, \theta) = \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1-x_i} = \theta \sum x_i (1 - \theta)^{1-\sum x_i}.$$ 

Any loss of information by using $\sum x_i$?

$$\begin{cases} 
\text{Yes} & \text{— can not examine the length of a run} \\
\text{No} & \text{— on inference of } \theta 
\end{cases}$$

**Heuristic:** Consider a vector of statistics $T(X)$, which summarizes the original data $X$. Then

- Full information, i.e. the information of $\theta$ contained in $X_1, X_2, \cdots X_n$
- $= \text{The information about } \theta \text{ given in } T(X) \text{ (reduced information)}$
- $+ \text{Given } T(X), \text{ the information of } \theta \text{ remained in } X_1, X_2, \cdots X_n \text{ (the rest information)}$.

**Definition.** A statistic is sufficient if given $T(X)$, the conditional distribution of $X$ is independent of $\theta$ — introduced by R.A.Fisher 1922.
Example 7 (continued). The conditional distribution of $X$ given $\sum_{i=1}^{n} X_i$ is

$$P_\theta \{ X = x | \sum_{i=1}^{n} X_i = s \} = \begin{cases} 0 & \text{if } \sum x_i \neq s, \\ \frac{P(X=x, \sum_{i=1}^{n} X_i = s)}{P(\sum_{i=1}^{n} X_i = s)} = \frac{\theta^s (1-\theta)^{n-s}}{\binom{n}{c} \theta^s (1-\theta)^{n-s}} & \text{otherwise}. \end{cases}$$

Obviously, this conditional distribution is independent of $\theta$. Thus, $\sum_{i=1}^{n} X_i$ is sufficient.

Theorem 1 (Factorization, Fisher-Neyman Theorem)

In a regular model, a statistic $T(X)$ is sufficient in $\theta \iff$

$$p(x, \theta) = g(T(x), \theta) h(x), \forall x \in \mathbb{R}^n \text{ and } \theta \in \Theta$$

for some functions $g(t, \theta)$ and $h$.

Proof: For simplicity to illustrate the idea, we concentrate on discrete case.
Suppose that $T(X)$ is sufficient. Then

$$p(x, \theta) = P_\theta[X = x, T(X) = T(x)]$$

$$= P_\theta[T(X) = T(x)]P_\theta[X = x | T(X) = T(x)]$$

$$= g(T(x), \theta)h(x).$$

Conversely,

$$P_\theta\{X = x | T(X) = T(x)\}$$

$$= \frac{P_\theta\{X = x\}}{P_\theta\{T(X) = T(x)\}}$$

$$= \frac{g(T(x), \theta)h(x)}{\sum\{y: T(y) = T(x)\} g(T(y), \theta)h(y)}$$

$$= \frac{h(x)}{\sum\{y: T(y) = T(x)\} h(y)}.$$

**Example 8.** Let $X_1, \cdots X_n$ be the inter-arrival times of $n$ customers with arrival rate $\theta$.

Then, under some conditions (rare; constant rate; independence) $X_1, X_2, \cdots X_n$
are i.i.d. random variables with \textit{Exponential}(\theta), i.e.

\[
p(X, \theta) = \prod_{i=1}^{n} \theta \exp(-\theta x_i) = \theta^n \exp(-\theta \sum_{i=1}^{n} x_i), \forall x_i \geq 0
\]

Hence, by taking \(g(t, \theta) = \theta^n \exp(-\theta t)\) and \(h(x) = 1\), we conclude that \(T(X) = \sum_{i=1}^{n} X_i\) is sufficient.

\textbf{Example 9.} (Size of population)
Then, $X_1, X_2, \cdots X_n$ are i.i.d. with

$$P(X_i = x_i) = \frac{1}{\theta} I\{1 \leq x_i \leq \theta\}.$$ 

Thus,

$$p(x, \theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} I\{1 \leq x_i \leq \theta\} = \theta^{-n} I\{\max\{x_i\} \leq \theta\},$$

and the largest order statistic $X_{(n)} = \max\{X_i\}$ is sufficient.

**Note:** This is not a realistic model. More realistic one is the capture-recapture model.

**Example 10** (Linear regression model). Suppose that $\{(X_i, Y_i)\}$ are a random sample from

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$
Then,
\[ p(X, y, \theta) \propto \prod_{i=1}^{n} \sigma^{-1} \exp\left(-\frac{1}{2\sigma^2}(Y_i - \alpha - \beta X_i)^2\right)f(X_i) \]
\[ = \prod_{i=1}^{n} f(X_i) \exp\left(-\log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} [Y_i - \alpha - \beta X_i]^2\right) \]
\[ \times \exp\left(-\frac{1}{2\sigma^2} \left[ \sum_{i=1}^{n} Y_i^2 - 2\alpha \sum_{i=1}^{n} Y_i - 2\beta \sum_{i=1}^{n} X_i Y_i \right] \right) \]
where \( f(\cdot) \) is density function of \( X \). Thus,
\[ T = \left( \sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} Y_i^2, \sum_{i=1}^{n} X_i Y_i, \sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2 \right) \]
is a sufficient statistic. This is equivalent to the fact that
\[ T^* = (\bar{X}, \bar{Y}, \hat{\sigma}_X^2, \hat{\sigma}_Y^2, r) \]
is a sufficient statistic.

**Sufficiency Principle**: Suppose that \( T(X) \) is sufficient. For any decision rule
\( \delta(X) \), we can find a decision rule \( \delta^*(T(X)) \), depending on \( T(X) \) and \( \delta(X) \) such that

\[
R(\theta, \delta) = R(\theta, \delta^*) \quad \text{for all} \quad \theta,
\]

where \( R(\theta, \delta) = E_\theta \ell(\theta, \delta(X)) \) is the expected loss function — risk function. Namely, considering the class of sufficient statistic is good enough for making statistical decisions.

**Proof.** For better understanding, let us first assume that \( \ell(\theta, a) \) is convex in \( a \). Then, let \( \delta^*(T) = E\{\delta(X)|T(X)\} \). By Jenssen’s inequality,

\[
E\ell(\theta, \delta(X)) = E\{E[\ell(\theta, \delta(X))|T]\}
\geq E\{\ell(\theta, \delta^*)\} = R(\theta, \delta^*).
\]

In general, let \( \delta^*(T(x)) \) be drawn at random from the conditional distribution \( \delta(x) \) given \( T(X) : \delta^* \sim L(\delta|T) \). Then,

\[
R(\theta, \delta) = E\{E[\ell(\theta, \delta)|T]\} = E\{E[\ell(\theta, \delta^*)|T]\} = R(\theta, \delta^*).
\]

**Sufficiency and Equivariant estimator**
**Example 11.** Suppose $X_1, X_2, \cdots, X_n \sim i.i.d. N(\mu, \sigma^2)$, e.g. measurement of temperature.

<table>
<thead>
<tr>
<th>data (in $^\circ C$)</th>
<th>data (in $^\circ F$/unnamed scale)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$ax_1 + b$</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$ax_2 + b$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$x_n$</td>
<td>$ax_n + b$</td>
</tr>
</tbody>
</table>

$\hat{\mu} : T(x_1, x_2, \cdots, x_n) \quad T(ax_1 + b, ax_2 + b, \cdots, ax_n + b)$

Estimate of $\mu$: $T(X_1, X_2, \cdots, X_n) \text{ in } ^\circ C = aT(X_1, X_2, \cdots, X_n) + b \text{ in } ^\circ F$

**Hope:** $T(ax_1 + b, ax_2 + b, \cdots, ax_n + b) = aT(x_1, x_2, \cdots, x_n) + b$

**Equivariance:** Such an estimator is called equivariant under linear transformation.

If we are interested in $\sigma$, we hope

$$T(X_1 + b, \cdots, X_n + b) = T(X_1, \cdots, X_n)$$
— invariant under the translation transform or more generally

\[ T(aX_1 + b, \cdots, aX_n + b) = aT(X_1, \cdots, X_n), \]

— equivariant under scale transformation /invariant under translations.

By sufficient principle, we need only to consider the estimator of form

\[ T(\bar{X}, S). \]

The equivariance for estimating \( \mu \) requires

\[ T(a\bar{X} + b, aS) = aT(\bar{X}, S) + b, \quad \forall a \text{ and } b \]

Taking \( a = 1 \) and \( b = -\bar{X} \), \( \implies T(0, S) = T(\bar{X}, S) - \bar{X} \)

\[ T(\bar{X}, S) = \bar{X} + T^*(S). \]

From

\[ T(a\bar{X}, aS) = a\bar{X} + T^*(aS) \]
\[ = a[\bar{X} + T^*(S)] \]
\[ \implies T^*(aS) = aT^*(S) \]
\[ \implies T^*(S) = ST^*(1). \]

Thus, denoting by \( T^* = T^*(1) \),

\[ T(\bar{X}, S) = \bar{X} + T^*S. \]

Among this invariant class,

\[ E[T(\bar{X}, S) - \mu]^2 = (ET^*S)^2 + \text{var}(\bar{X} + T^*S) \]
\[ = T^{*2}(ES)^2 + T^{*2}\text{var}(S) + \sigma^2/n \]
It attains the minimum at \( T^* = 0 \), namely, \( \bar{X} \) is the best equivalent estimator.

### Sufficiency and Bayesian Model

**Theorem 2 (Kolmogrov)** If \( T(X) \) is sufficient for \( \theta \), then for any prior \( \pi(\theta) \), the conditional distribution

\[
\mathcal{L}(\theta|T(X)) = \mathcal{L}(\theta|X) — \text{Bayes sufficient.}
\]

According to the theorem,

\[
E(g(\theta)|T) = E(g(\theta)|X).
\]

This implies that given \( T(X) \), and \( X \) and \( \theta \) are independent, since

\[
E[f(\theta)g(X)|T] = E[E(f(\theta)g(X)|X)|T]
\]

\[
= E[g(X)E(f(\theta)|T)|T]
\]

\[
= E[g(X)|T]E[f(\theta)|T].
\]
1.4 Exponential Families

Many useful distributions admit a common structure:

Normal (continuous), Poisson (counts)

**Examples**  Binomial (categorical), Beta

Gamma (constant Coefficient of Variation)

They form the basis of GLIM (Generalized Linear Models). Such a family is called exponential families, discovered independently by Koopman, Pitman and Darmois. It is nice to give them a unified mathematical treatment.

**The one parameter case**

**Example 12.** Let $P_\theta = \{N(\mu, \sigma_0^2), \sigma_0 \text{ is known}\}$. Then its density

$$
p(x, \mu) = \frac{1}{\sqrt{2\pi\sigma_0}} \exp \left( -\frac{(x - \mu)^2}{2\sigma_0^2} \right)
= \exp \left\{ \frac{x\mu}{\sigma_0^2} - \frac{\mu^2}{2\sigma_0^2} - \left( \frac{x^2}{2\sigma_0^2} + \log \sqrt{2\pi\sigma_0} \right) \right\}
= \exp \left( T(x)c(\theta) + d(\theta) + S(x) \right).
$$
**Example 13.** Let $P_\theta = \{\text{Binomial}(n, \theta)\}$. Then,

$$p(x, \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$$

$$= \exp \left\{ x \log \frac{\theta}{1 - \theta} + n \log(1 - \theta) + \log \binom{n}{x} \right\}$$

$$= \exp \left\{ T(x)c(\theta) + d(\theta) + S(x) \right\} .$$

**Definition:** The family of distributions of a model $\{P_\theta : \theta \in \Theta\}$ is said to be a one-parameter exponential one if

$$p(x, \theta) = \exp\{c(\theta)T(x) + d(\theta) + S(x)\} .$$

**Example 14.** Let $X \sim \text{Unif}(0, \theta)$. Then

$$p(x, \theta) = \frac{1}{\theta} I_{[0,\theta]}(x) = \exp(\log I_{[0,\theta]}(x) - \log \theta) ,$$

not an exponential family. Another example is

$$p(x, \theta) = \frac{1}{9} I(x \in \{0.1 + \theta, \cdots , 0.9 + \theta\}) .$$
By setting $c(\theta) = \eta$, the exponential family can be written in the canonical form as

$$p(x, \eta) = \exp(\eta T(x) + d_0(\eta) + S(x)),$$

where $d_0(\eta) = d(c^{-1}(\eta))$, when $c(\theta)$ is one-to-one.

$\eta$ — canonical (natural) parameter and $c(\cdot)$ — canonical link,

**Examples** of canonical link functions:

- Normal $c(\theta) = \theta$ identity
- Binomial $c(\theta) = \log \frac{\theta}{1-\theta}$ logit
- Poisson $c(\theta) = \log \theta$ logarithm.

**Regeneration properties:**

1. Let $X_1, \cdots, X_n \sim i.i.d. P_\theta$, belonging to an exponential family. Then, the joint density $\prod_{i=1}^n p(x_i, \theta)$ is also in the exponential family. Further, $\sum_{i=1}^n T(X_i)$ is a sufficient statistic.
2. If $X \sim P_\theta$ which is exponential family, and $\{Q_\theta\}$ be the distribution of $T(X)$, Then, $\{Q_\theta\}$ is also in the exponential family.

**Theorem 3** If $X \sim \exp\{\eta T(X) + d_0(\eta) + S(x)\}$, $\eta$ is an interior of $\mathcal{E}$, then

$$\psi(s) = E \exp\{sT(X)\} = \exp[d_0(\eta) - d_0(s + \eta)], \text{ for } s \text{ near } 0$$

Moreover, $ET(X) = -d'_0(\eta)$, $\text{var}(T(x)) = -d''_0(\eta)$. (The function $d_0$ is concave.)

**Proof:** Note that

$$\int_{-\infty}^{+\infty} \exp\{\eta T(x) + d_0(\eta) + S(x)\} \, dx = 1,$$

$$\implies \int_{-\infty}^{+\infty} \exp\{\eta T(x) + S(x)\} \, dx = \exp (-d_0(\eta)).$$
Now,

\[ \psi(s) = E\{\exp(sT(x))\} \]

\[ = \int_{-\infty}^{+\infty} \exp\{sT(x) + \eta T(x) + d_0(\eta) + S(x)\} \, dx \]

\[ = \exp(d_0(\eta) - d_0(\eta + s)). \]

From the properties of the moment generating function,

\[ \psi'(s)|_{s=0} = E\{T(X) \exp(sT(X))|_{s=0}\} \]

\[ = ET(X) \]

\[ = -\exp(d_0(\eta) - d_0(\eta + s))d'_0(\eta + s)|_{s=0}. \]

Similarly,

\[ ET^2(X) = \psi''(s)|_{s=0} = -d''_0(\eta) + d'_0(\eta)^2 \]

\[ \implies \text{var}(T(X)) = -d''_0(\eta). \]
**Example 15.** \( X_1, \cdots, X_n \sim \text{i.i.d.} \)

\[
p(x, \theta) = k\theta (\theta x)^{k-1} \exp(- (\theta x)^k), \ x > 0.
\]

— Weibull distribution \( \Rightarrow \) model “failure time” with hazard risk: \( \frac{f(t)}{1-F(t)} = k\theta (\theta t)^{k-1} \)

\( k = 1 \) \( \Rightarrow \) exponential distribution — constant risk

\( k = 2 \) \( \Rightarrow \) Raleigh distribution — \( k\theta^2 t \) (linear risk)

Then, the joint density

\[
p(x, \theta) = \prod_{i=1}^{n} k\theta (\theta x_i)^{k-1} \exp(- \theta^k x_i^k)
\]

\[
= \exp(- \theta^k \sum_{i=1}^{n} x_i^k - nk \log \theta + \sum_{i=1}^{n} \log x_i^{k-1} + n \log k).
\]

For this family of distribution, \( \eta = -\theta^k \)

\[
d_0(\eta) = -n \log \theta^k = -n \log(-\eta).
\]
Hence,

\[
\sum_{i=1}^{n} X_i^k \quad \text{— natural sufficient statistic,}
\]

\[
E \sum_{i=1}^{n} X_i^k = \frac{-n}{\eta} = \frac{n}{\theta^k},
\]

\[
\text{var}(\sum_{i=1}^{n} X_i^k) = \frac{n}{\eta^2} = \frac{n}{\theta^{2k}}.
\]

Direct computation of these moments are more complicated.

**The \( k \) parameter case**

A family of distributions \( \{P_\theta : \theta \in \Theta\} \) is said to be \( k \) parameter exponential family if its joint density admits the form

\[
p(x, \theta) = \exp\left(\sum_{i=1}^{k} C_i(\theta)T_i(x) + d(\theta) + S(x)\right)
\]

\[
= \exp\left(\sum_{i=1}^{k} \eta_i T_i(x) + d_0(\eta)\right).
\]
By the factorization theorem, the vector \( T(x) = (T_1(x), \cdots, T_k(x)) \) is a sufficient statistic.

Suppose that \( \mathbf{X}_1, \cdots, \mathbf{X}_n \) are a random sample from \( P_{\theta} \). Put \( \mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n) \) which is available data.

Then, the distribution of \( \mathbf{X} \) forms a \( k \)-parametric family with

\[
T(\mathbf{X}) = \left( \sum_{i=1}^{n} T_1(\mathbf{X}_i), \cdots, \sum_{i=1}^{n} T_k(\mathbf{X}_i) \right)
\]

Let \( \psi(s) = E \exp(s^T T(x)) \). Then,

\[
\psi(s) = \exp(d_0(\eta) - d_0(\eta + s))
\]

\[
ET(x) = -d_0'(\eta) — \text{mean vector}
\]

\[
\text{var}(T(x)) = -d_0''(\eta) — \text{variance-covariance matrix}
\]

**Example 16.** (Multinomial trails)

\[
P(X_i = j) = p_j = \prod_{\ell=1}^{k} p^{I(j=\ell)}
\]
Figure 1.12: Multinomial trial. Each outcome is a $k$-dimensional unit vector, indicating which category is observed.

\[
\prod_{i=1}^{n} P(x_i, p) = \prod_{\ell=1}^{k} \prod_{i=1}^{n} P^{I(x_i=\ell)} = \prod_{\ell=1}^{k} p_{\ell}^{n_{\ell}}.
\]

\[
n_{\ell} = \sum_{i=1}^{n} I(x_i = \ell) \quad \# \text{ of times observing } \ell
\]

The joint density is

\[
p(x, p) = \exp\left\{ \sum_{\ell=1}^{k} n_{\ell} \log p_{\ell} \right\}
\]

\[
= \exp\left\{ \sum_{\ell=1}^{k-1} n_{\ell} \log \frac{p_{\ell}}{p_{k}} + n \log p_{k} \right\}.
\]
Let $\alpha_j = \log p_j - \log p_k, j = 1, \cdots, k - 1$. Then

$$p_k = 1 - p_1 - \cdots - p_{k-1} = 1 - p_k \sum_{j=1}^{k-1} e^{\alpha_j}$$

$$\implies p_k = \frac{1}{1 + \sum_{j=1}^{k-1} e^{\alpha_j}}$$

Hence,

$$p(x, p) = \exp\left\{ \sum_{\ell=1}^{k-1} n_\ell \alpha_\ell - n \log(1 + \sum_{j=1}^{k-1} e^{\alpha_j}) \right\}.$$  

The variance and covariance matrix of $(n_1, \cdots, n_k)$ can easily be completed.

**Other Examples** — Multivariate normal distributions

— Dirichlet distribution (multivariate $\beta$-distribution):

$$cx_1^{\beta_1-1} \cdots x_p^{\beta_p-1}(1 - x_1 - \cdots - x_p)^{\beta_p+1-1}.$$