Chapter 1

Statistical Modeling

1.1 Statistical Models

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

Possible outcomes: GGDGGGDD $\cdots$, realization of outcomes.

How do we connect the sample with the population?

Modelling — think of data as a realization of a the random experiment.

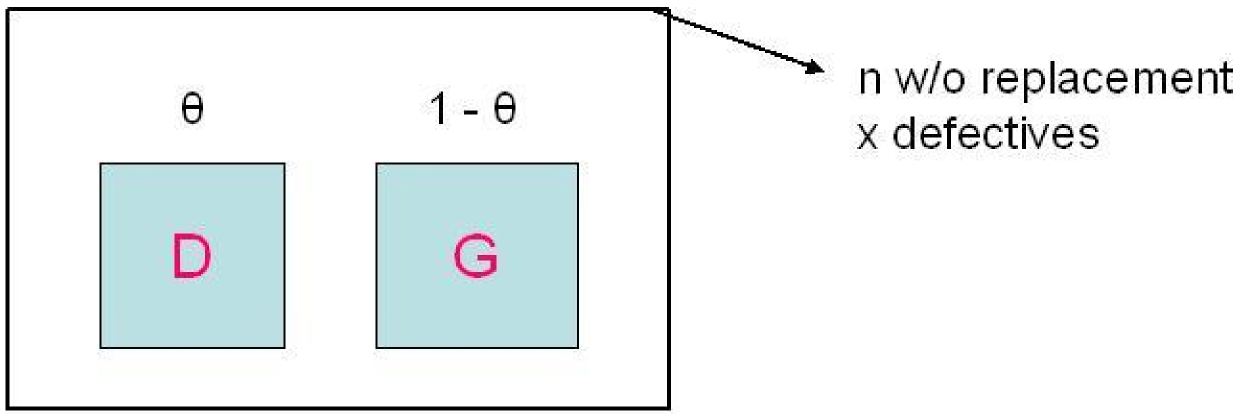

Figure 1.1: Illustration of the sampling scheme.

Observe that a "D" $\Longrightarrow \theta$ is large,

a "G" $\Longrightarrow \theta$ 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 $\max(0, n - N(1-\theta)) \leqslant x \leqslant \min(n, N\theta)$. 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: θ — unknown, fixed.

Parameter space Θ: the possible value of θ: $\Theta = \{0/N, 1/N, \cdots, N/N\}$ 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 θ?

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 θ 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 μ 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 ε_i is error of measurement noise.

Assumptions

i) ε_i is independent of μ.

ii) $\varepsilon_i, i = 1, 2, \cdots, n$ are independent.

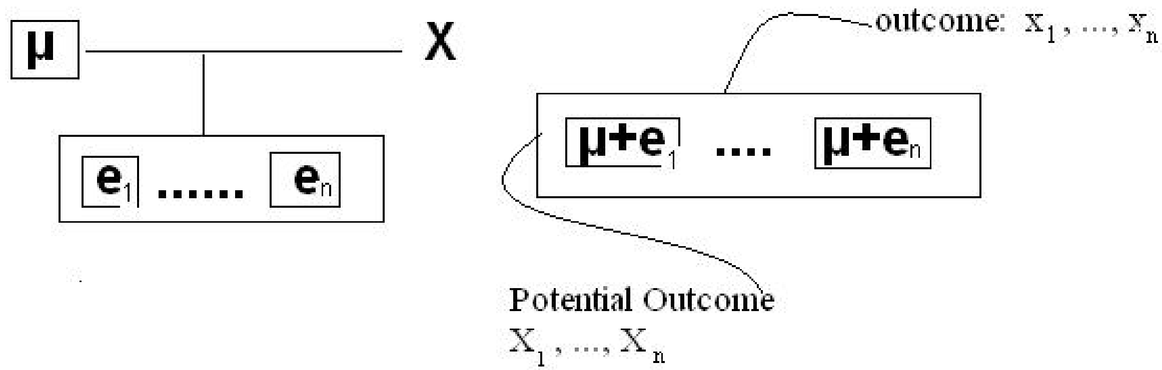

Figure 1.2: Illustration of the idea of modeling.

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

iv) the distribution of ε 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 σ^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 μ?

Logically, if $\mu = 100$, it is possible to observe $\mathbf{x}$. If $\mu = 1$, it is also possible to observe $\mathbf{x}$. So we can not absolutely tell what value of μ is. But from the square-root law:

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

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

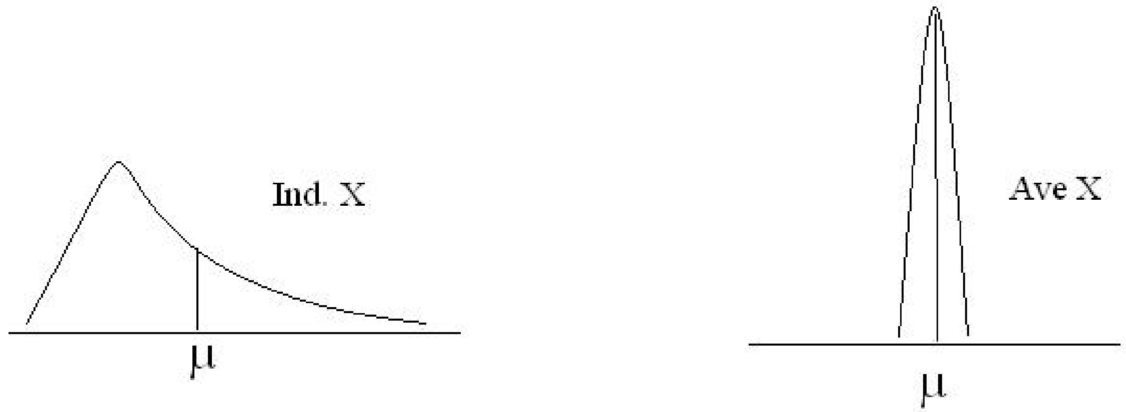

Figure 1.3: Distributions of individual observation versus that of average

Example 3: Drug evaluation (Hypertension drug)

Drug A $\longrightarrow$ m patiets Drug B $\longrightarrow$ n patiets

Measurement: blood pressure.

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

Drug A					Drug B					
150	110	160	187	153	120	140	160	180	133	136
x_1	x_2	x_3	x_4	x_5	y_1	y_2	y_3	y_4	y_5	y_6

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

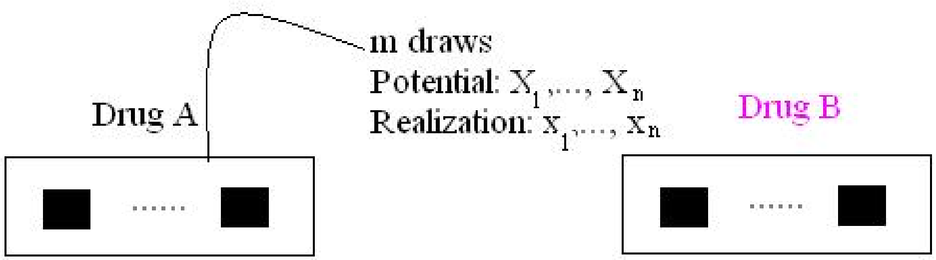

Figure 1.4: Illustration of a two-sample problem

	Drug A	Drug B
random outcomes	$X_1, \cdots, X_m$	$Y_1 \cdots, Y_n$
realizations	$x_1, \cdots, x_m$	$y_1, \cdots, y_n$

Further, we might assume that

$$X_1, \cdots, X_m \overset{i.i.d}{\sim} N(\mu_A, \sigma_A^2) \qquad Y_1, \cdots, Y_n \overset{i.i.d}{\sim} N(\mu_B, \sigma_B^2).$$

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 σ.

Connection sample with population: data are realizations from a population, whose distribution depends on θ.

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 $\mathbf{X}$ is assumed in $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, Θ is the parametric space.

Objectives: Inferences about θ.

— In Example 1:

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

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

— In Example 2:

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

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

— In Example 3:

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

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 θ doesn't have to be in $\mathbb{R}^k$. In Example 2, without the normality assumption,

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

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 Θ is called nonparametric or semiparametric.

Basic assumption: Throughout this class, we will assume that

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

(ii) Discrete variable:All P_θ 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 θ.

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

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

$$X_1, \cdots, X_m \overset{i.i.d}{\sim} N(\mu + \alpha_1, \sigma^2) \qquad Y_1, \cdots, Y_n \overset{i.i.d}{\sim} N(\mu + \alpha_2, \sigma^2).$$

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

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

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 θ 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

$\{(x_{i1}, \cdots, x_{ip}, y_i)\}_{i=1}^n$ 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

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

where ε 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 $\mathbf{X}$ 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.

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

but

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

are not.

Estimator: an estimating procedure for certain parameters, e.g. $\overline{X}$ for μ.

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

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

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:

$$\begin{cases} \text{long run relative frequency — Frequentist} \\ \text{prior knowledge w/brief — Bayesian} \end{cases}$$

So far, we have assumed no information about θ beyond that provided by data. Often, we can have some (vague) knowledge about θ. 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 $\pi(\theta)$:

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

This provides as a prior distribution. The defective rate θ_0 of the current lot is thought of as a realization from $\pi(\theta)$. Given θ_0,

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

Basic element of Baysian models

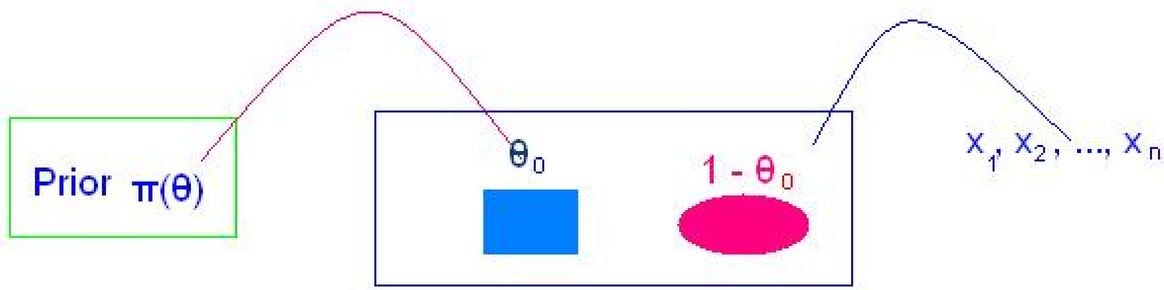

Figure 1.5: Bayesian Framework

(i) The knowledge about θ is summarized by $\pi(\theta)$ — prior dist.

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

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

(iv) The goal of the Bayesian analysis is to modify the prior of θ after observing $\mathbf{x}$:

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

e.g. summarizing the distribution by posterior mean, median and SD, etc.

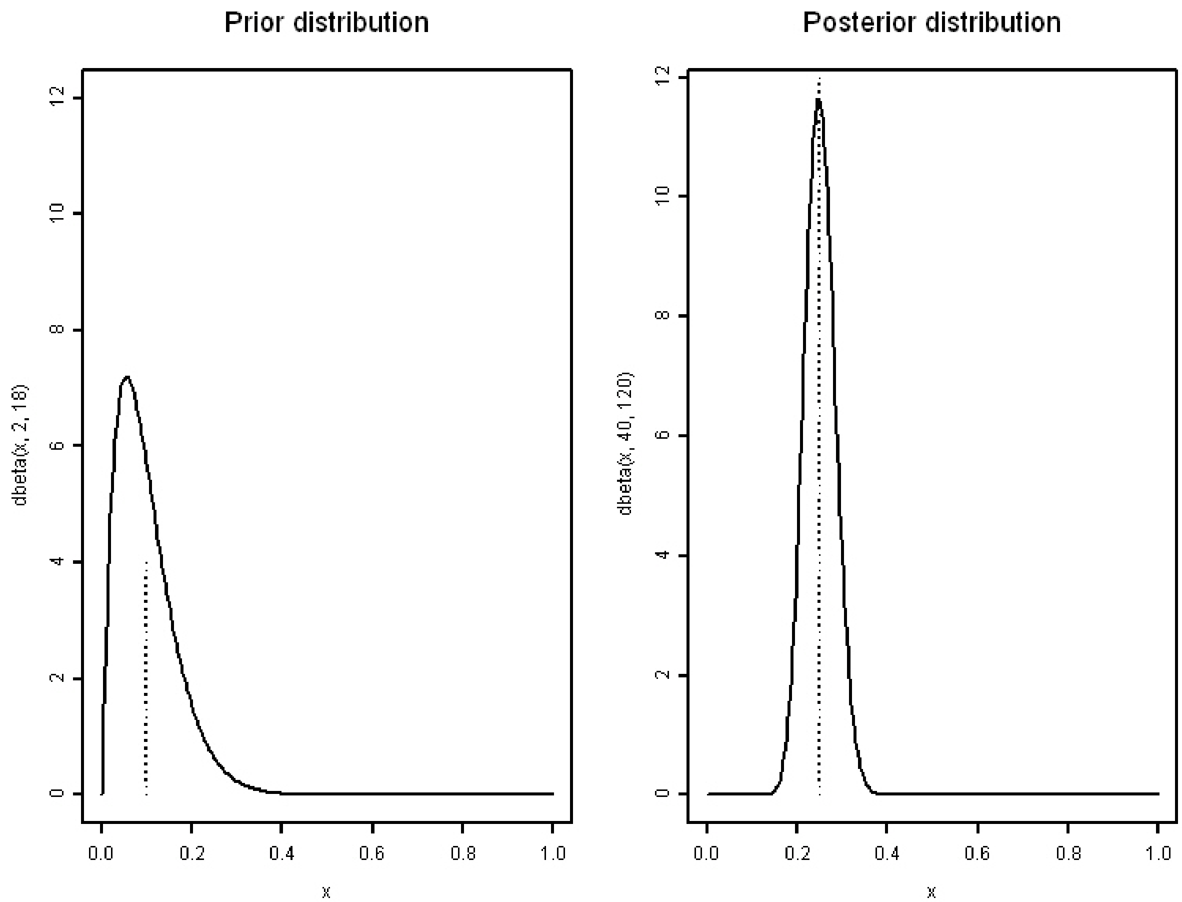

Figure 1.6: Prior versus Posterior distributions

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.

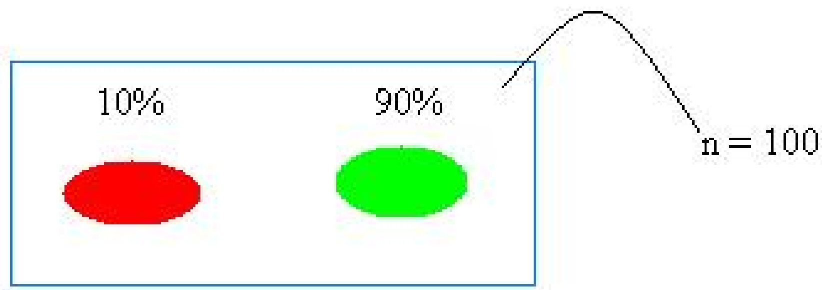

Figure 1.7: Prior knowledge of the defects

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

$$E\theta = E\frac{X}{100} = 0.1$$

$$\text{var}(\theta) = \frac{1}{100^2}\text{var}(X) = \frac{100\times0.9\times0.1}{100^2},$$

$$SD(\theta) = 0.03.$$

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.

$$\begin{aligned} P(\theta \geqslant 0.2|X=10) &= P(100\theta - X \geqslant 10|X=10) \\ &\approx 1-\Phi\left(\frac{10-81\times 0.1}{\sqrt{81\times 0.9\times 0.1}}\right) \\ &\approx 30\%. \end{aligned}$$

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

$$\begin{aligned} P(\theta \geqslant 0.2) &= P(100\theta \geqslant 20) \\ &= 1-\Phi\left(\frac{20-100\times 0.1}{\sqrt{100\times 0.9\times 0.1}}\right) \\ &\approx 0.1\%, \end{aligned}$$

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(θ) and θ has a prior distribution $\pi(\theta)$. Then

$$\pi(\theta|\mathbf{x}) = \frac{\pi(\theta)\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}{\int_0^1 \pi(t)t^{\sum_{i=1}^n x_i}(1-t)^{n-\sum_{i=1}^n x_i}dt}.$$

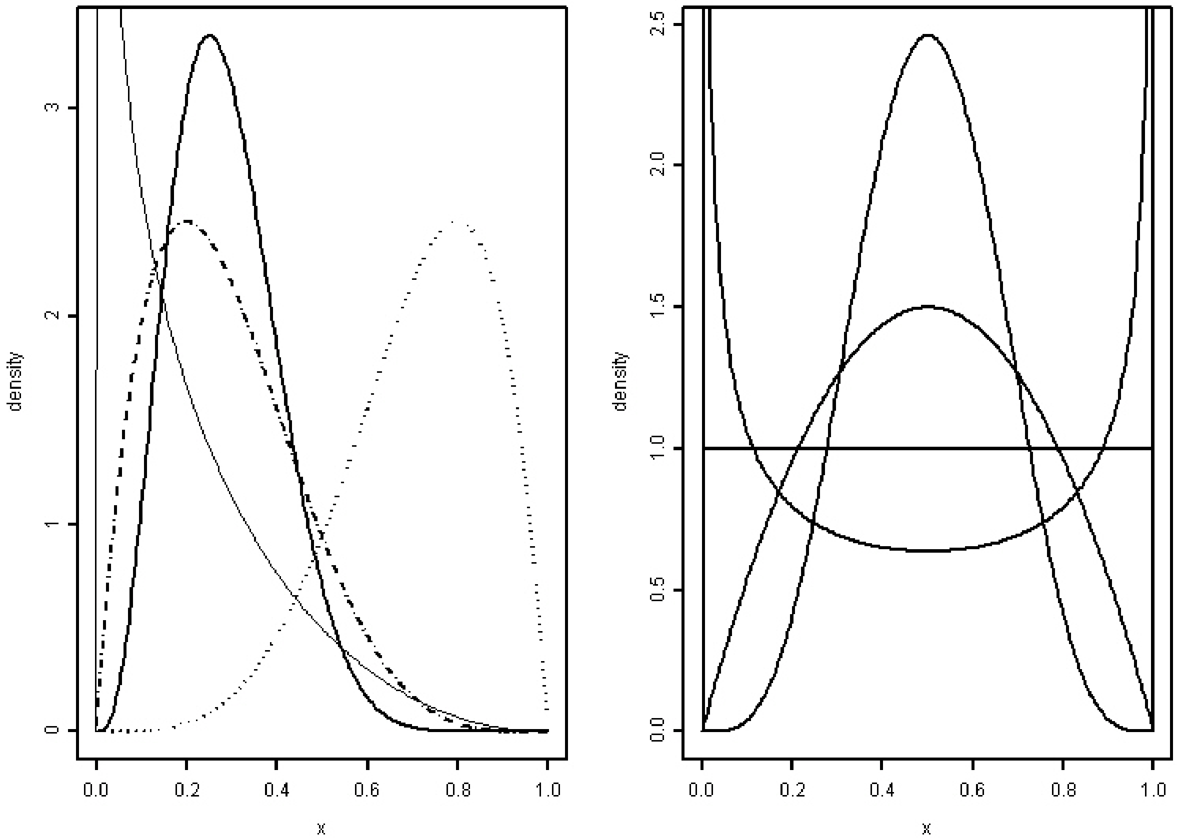

Figure 1.8: Beta distributions with shape parameters: Left panel: (4, 10), (5, 2), (2, 5), (.7, 3); right panel: (5, 5), (2, 2), (1, 1), (0.5, 0.5)

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

$$\pi(\theta) = \frac{\theta^{s-1}(1-\theta)^{r-1}}{B(s,t)}, \quad E\theta = \frac{s}{r+s},$$

then

$$\pi(\theta|\mathbf{x}) \propto \theta^{s+\sum x_i - 1}(1-\theta)^{n-\sum x_i + r} \sim Beta(s + \sum x_i, n - \sum x_i + r).$$

Thus,

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

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

$$\begin{cases} 1^o \text{ Sufficiency} \\ 2^o \text{ Invariant/equivariant} \end{cases}$$

Purpose:

$$\begin{cases} 1 \text{ simplify probability structure, less obscure than the whole data} \\ 2 \text{ understand whether a loss in reduction} \\ 3 \text{ useful technical tools} \end{cases}$$

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

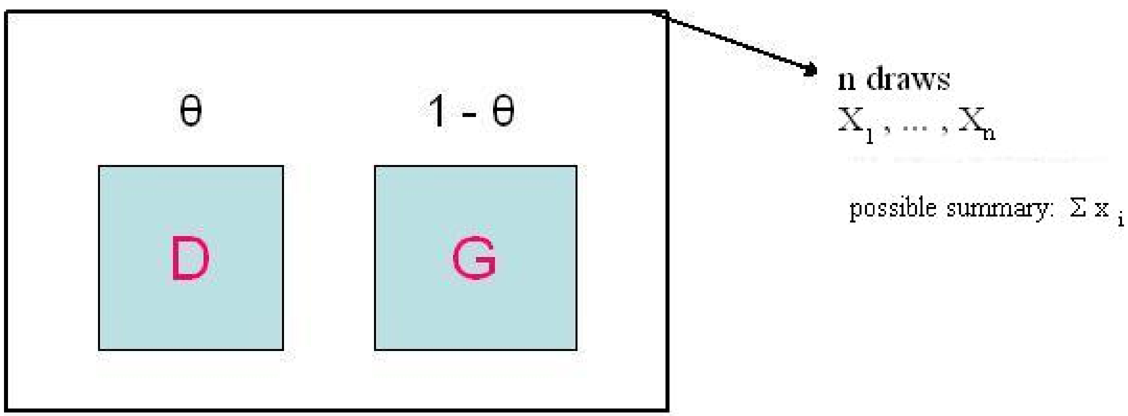

Figure 1.9: Probability model and its summary statistic.

Then, the probability model is

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

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 $T(\mathbf{X})$, which summarizes the original data $\mathbf{X}$. Then

Full information, i.e. the information of θ contained in $X_1, X_2, \cdots X_n$

= The information about θ given in $T(\mathbf{X})$(reduced information)

+ Given $T(\mathbf{X})$, the information of θ remained in $X_1, X_2, \cdots X_n$(the rest information).

Definition. A statistic is sufficient if given $T(\mathbf{X})$, the conditional distribution of $\mathbf{X}$ is independent of θ — introduced by R.A.Fisher 1922.

Example 7 (continued). The conditional distribution of $\mathbf{X}$ given $\sum_{i=1}^n X_i$ is

$$P_\theta\{\mathbf{X} = \mathbf{x} | \sum_{i=1}^n X_i = s\}$$

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

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

Theorem 1 *(*Factorization, Fisher-Neyman Theorem*)*

In a regular model, a statistic $T(\mathbf{X})$ is sufficient in $\theta \Longleftrightarrow$

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

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

Proof: For simplicity to illustrate the idea, we concentrate on discrete case.

Suppose that $T(\mathbf{X})$ is sufficient. Then

$$\begin{aligned} p(\mathbf{x}, \theta) &= P_\theta[\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = T(\mathbf{x})] \\ &= P_\theta[T(\mathbf{X}) = T(\mathbf{x})] P_\theta[\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = T(\mathbf{x})] \\ &= g(T(\mathbf{x}), \theta) h(\mathbf{x}). \end{aligned}$$

Conversely,

$$\begin{aligned} & P_\theta\{\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = T(\mathbf{x})\} \\ = \; & \frac{P_\theta\{\mathbf{X} = \mathbf{x}\}}{P_\theta\{T(\mathbf{X}) = T(\mathbf{x})\}} \\ = \; & \frac{g(T(\mathbf{x}), \theta) h(\mathbf{x})}{\sum_{\{y: T(\mathbf{y}) = T(\mathbf{x})\}} g(T(\mathbf{y}), \theta) h(\mathbf{y})} \\ = \; & \frac{h(\mathbf{x})}{\sum_{\{y: T(\mathbf{y}) = T(\mathbf{x})\}} h(\mathbf{y})}. \end{aligned}$$

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

Then, under some conditions (rare; constant rate; independence) $X_1, X_2, \cdots X_n$

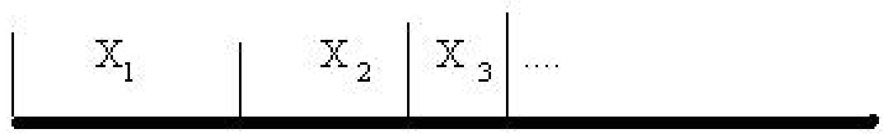

Figure 1.10: Arrival times of customer

are $i.i.d.$ random variables with $Exponential(\theta)$, i.e.

$$p(\mathbf{X}, \theta) = \Pi_{i=1}^n \theta \exp(-\theta x_i) = \theta^n \exp(-\theta \sum_{i=1}^n x_i), \forall x_i \geqslant 0$$

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

Example 9.(Size of population)

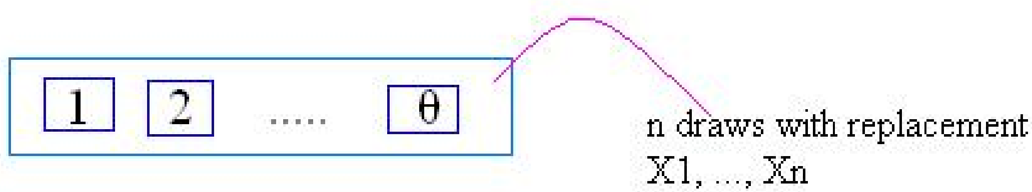

Figure 1.11: Estimation the size of population

Then, $X_1, X_2, \cdots X_n$ are i.i.d. with

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

Thus,

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

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

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2).$$

Then,

$$
\begin{aligned}
& p(\mathbf{X}, \mathbf{y}, \theta) \\
\propto\ & \Pi_{i=1}^n \sigma^{-1} \exp(-\frac{1}{2\sigma^2}(Y_i - \alpha - \beta X_i)^2) f(X_i) \\
=\ & \Pi_{i=1}^n f(X_i) \exp\left(-\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n [Y_i - \alpha - \beta X_i]^2\right) \\
& \times \exp\left(-\frac{1}{2\sigma^2}[\sum_{i=1}^n Y_i^2 - 2\alpha \sum_{i=1}^n Y_i - 2\beta \sum_{i=1}^n X_i Y_i]\right)
\end{aligned}
$$

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

$$
T^* = (\bar{X}, \bar{Y}, \widehat{\sigma}_X^2, \widehat{\sigma}_Y^2, r)
$$

is a sufficient statistic.

Sufficiency Principle: Suppose that $T(\mathbf{X})$ is sufficient. For any decision rule

$\delta(\mathbf{X})$, we can find a decision rule $\delta^*(T(\mathbf{X}))$, depending on $T(\mathbf{X})$ and $\delta(\mathbf{X})$ such that

$$R(\theta, \delta) = R(\theta, \delta^*) \text{ for all } \theta,$$

where $R(\theta, \delta) = E_\theta \ell(\theta, \delta(\mathbf{X}))$ 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 $\delta^*(T) = E\{\delta(\mathbf{X})|T(\mathbf{X})\}$. By Jenssen's inequality,

$$\begin{aligned} E\ell(\theta, \delta(\mathbf{X})) &= E\{E[\ell(\theta, \delta(\mathbf{X}))|T]\} \\ &\geq E\{\ell(\theta, \delta^*)\} = R(\theta, \delta^*). \end{aligned}$$

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

data (in ^oC)	data(in ^oF/unnamed scale)
x_1	$ax_1 + b$
x_2	$ax_2 + b$
$\vdots$	$\vdots$
x_n	$ax_n + b$
$\widehat{\mu}$: $T(x_1, x_2, \cdots, x_n)$	$T(ax_1 + b, ax_2 + b, \cdots, ax_n + b)$

Estimate of μ: $T(X_1, X_2, \cdots, X_n)$ in $^oC = aT(X_1, X_2, \cdots, X_n) + b$ in ^oF

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 σ, 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 μ requires

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

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

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

From

$$\begin{aligned} T(a\bar{X}, aS) &= a\bar{X} + T^*(aS) \\ &= a[\bar{X} + T^*(S)] \\ &\Longrightarrow T^*(aS) = aT^*(S) \\ &\Longrightarrow T^*(S) = ST^*(1). \end{aligned}$$

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

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

Among this invariant class,

$$\begin{aligned} E[T(\bar{X}, S) - \mu]^2 &= (ET^* S)^2 + \text{var}(\bar{X} + T^* S) \\ &= T^{*2}(ES)^2 + T^{*2}\text{var}(S) + \sigma^2/n \end{aligned}$$

It attains the minimum at $T^* = 0$, namely, $\bar{X}$ is the best equivalent estimator.

Sufficiency and Bayesian Model

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

$$\mathcal{L}(\theta|T(\mathbf{X})) = \mathcal{L}(\theta|\mathbf{X})\text{—}\textit{Bayes sufficient.}$$

According to the theorem,

$$E(g(\theta)|T) = E(g(\theta)|\mathbf{X}).$$

This implies that given $T(\mathbf{X})$, and $\mathbf{X}$ and θ are independent, since

$$\begin{aligned} E[f(\theta)g(\mathbf{X})|T] &= E[E(f(\theta)g(\mathbf{X})|\mathbf{X})|T] \\ &= E[g(\mathbf{X})E(f(\theta)|T)|T] \\ &= E[g(\mathbf{X})|T]E[f(\theta)|T]. \end{aligned}$$

1.4 Exponential Families

Many useful distributions admit a common structure:

Examples
Normal (continuous), Poisson (counts)
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

$$\begin{aligned} p(x,\mu) &= \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{(x-\mu)^2}{2\sigma_0^2}\right) \\ &= \exp\left\{\frac{x\mu}{\sigma_0^2} - \frac{\mu^2}{2\sigma_0^2} - \left(\frac{x^2}{2\sigma_0^2} + \log\sqrt{2\pi}\sigma_0\right)\right\} \\ &= \exp\left(T(x)c(\theta) + d(\theta) + S(x)\right). \end{aligned}$$

Example 13. Let $P_\theta = \{Binomial(n, \theta)\}$. Then,

$$\begin{aligned} p(x, \theta) &= \binom{n}{x} \theta^x (1-\theta)^{n-x} \\ &= \exp\left\{ x \log \frac{\theta}{1-\theta} + n \log(1-\theta) + \log \binom{n}{x} \right\} \\ &= \exp\left\{ T(x) c(\theta) + d(\theta) + S(x) \right\}. \end{aligned}$$

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.

η — 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 $\Pi_{i=1}^n p(x_i, \theta)$ 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* $X \sim \exp\{\eta T(X) + d_0(\eta) + S(x)\}$, η *is an interior of* $\mathcal{E}$, *then*

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

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

Proof: Note that

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

$$\Longrightarrow \quad \int_{-\infty}^{+\infty} \exp\{\eta T(x) + S(x)\}\, dx = \exp\left(-d_0(\eta)\right).$$

Now,

$$\begin{aligned}
\psi(s) &= E\{\exp(sT(x))\} \\
&= \int_{-\infty}^{+\infty} \exp\{sT(x) + \eta T(x) + d_0(\eta) + S(x)\}\, dx \\
&= \exp(d_0(\eta) - d_0(\eta + s)).
\end{aligned}$$

From the properties of the moment generating function,

$$\begin{aligned}
\psi'(s)|_{s=0} &= E\{T(X)\exp(sT(X))|_{s=0}\} \\
&= ET(X) \\
&= -\exp(d_0(\eta) - d_0(\eta + s))d_0'(\eta + s)|_{s=0}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
& ET^2(X) = \psi''(s)|_{s=0} = -d_0''(\eta) + d_0'(\eta)^2 \\
\Longrightarrow \quad & \mathrm{var}(T(X)) = -d_0''(\eta).
\end{aligned}$$

Example 15. $X_1, \cdots, X_n \sim i.i.d.$

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

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

$k = 1 \Longrightarrow$ exponential distribution — constant risk

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

Then, the joint density

$$\begin{aligned} p(\mathbf{x}, \theta) &= \Pi_{i=1}^n k\theta(\theta x_i)^{k-1} \exp(-\theta^k x_i^k) \\ &= \exp(-\theta^k \sum_{i=1}^n x_i^k - nk \log \theta + \sum_{i=1}^n \log x_i^{k-1} + n \log k). \end{aligned}$$

For this family of distributionm,

$$\eta = -\theta^k$$

$$d_0(\eta) = -n \log \theta^k = -n \log(-\eta).$$

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

$$\begin{aligned} p(\mathbf{x}, \theta) &= \exp(\sum_{i=1}^{k} C_i(\theta) T_i(\mathbf{x}) + d(\theta) + S(\mathbf{x})) \\ &= \exp(\sum_{i=1}^{k} \eta_i T_i(\mathbf{x}) + d_0(\eta)). \end{aligned}$$

By the factorization theorem, the vector $T(\mathbf{x}) = (T_1(\mathbf{x}), \cdots, T_k(\mathbf{x}))$ is a sufficient statistic.

Suppose that $\mathbf{X}_1, \cdots, \mathbf{X}_n$ are a random sample from P_θ. 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

$$T(\mathbf{X}) = (\sum_{i=1}^n T_1(\mathbf{X}_i), \cdots, \sum_{i=1}^n T_k(\mathbf{X}_i))$$

Let $\psi(\mathbf{s}) = E\exp(\mathbf{s}^T T(\mathbf{x}))$. Then,

$$\psi(s) = \exp(d_0(\eta) - d_0(\eta + \mathbf{s}))$$

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

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

Example 16. (Multinomial trails)

$$P(X_i = j) = p_j = \Pi_{\ell=1}^k p_\ell^{I(j=\ell)}$$

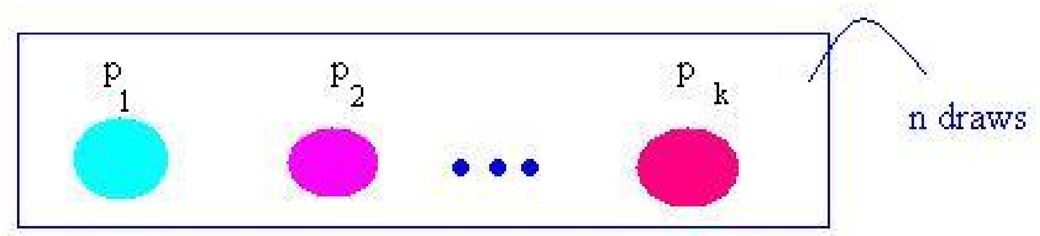

Figure 1.12: Multinomial trial. Each outcome is a k-dimensional unit vector, indicting which category is observed.

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

$$n_\ell = \sum_{i=1}^n I(x_i = \ell) \text{ — } \sharp \text{ of times observing } \ell$$

The joint density is

$$\begin{aligned} p(\mathbf{x}, \mathbf{p}) &= \exp\{\sum_{\ell=1}^k n_\ell \log p_\ell\} \\ &= \exp\{\sum_{\ell=1}^{k-1} n_\ell \log \frac{p_\ell}{p_k} + n \log p_k\}. \end{aligned}$$

Let $\alpha_j = \log p_j - \log p_k, j = 1, \cdots, k-1$. Then

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

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

Hence,

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

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

Other Examples: — Multivariate normal distributions

— Dirichlet distribution (multivariate β-distribution):

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