Lecture 17: Populations, samples, models, and statistics

One or a series of random experiments is performed.
Some data from the experiment(s) are collected.
Planning experiments and collecting data (not discussed in the textbook).
Data analysis: extract information from the data, interpret the results, and draw some conclusions.

A descriptive data analysis: summary measures of the data, such as the mean, median, range, standard deviation, etc., and some graphical displays, such as the histogram and box-and-whisker diagram, etc.

It is simple and requires almost no assumptions, but may not allow us to gain enough insight into the problem.
We focus on more sophisticated methods of analyzing data: *statistical inference* and *decision theory*.

The data set is a realization of a random element defined on a probability space $(\Omega, \mathcal{F}, P)$
P is called the *population*.
The data set or the random element that produces the data is called a *sample* from P.
The size of the data set is called the *sample size*.
A population P is *known* if and only if $P(A)$ is a known value for every event $A \in \mathcal{F}$.
In a statistical problem, the population P is at least partially unknown.
We would like to deduce some properties of P based on the available sample.

Examples 2.1-2.3

A *statistical model* (a set of assumptions) on the population P in a given problem is often postulated to make the analysis possible or easy.
Although testing the correctness of postulated models is part of statistical inference and decision theory, postulated models are often based on knowledge of the problem under consideration.

Definition 2.1. A set of probability measures P_θ on $(\Omega, \mathcal{F})$ indexed by a *parameter* $\theta \in \Theta$ is said to be a *parametric family* if and only if $\Theta \subset \mathcal{R}^d$ for some fixed positive integer d *and* each P_θ is a *known* probability measure when θ is known. The set Θ is called the *parameter space* and d is called its *dimension*.

Parametric model: the population P is in a parametric family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$
$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is *identifiable* if and only if $\theta_1 \neq \theta_2$ and $\theta_i \in \Theta$ imply $P_{\theta_1} \neq P_{\theta_2}$.
In most cases an identifiable parametric family can be obtained through reparameterization.

A family of populations $\mathcal{P}$ is dominated by ν (a σ-finite measure) if $P \ll \nu$ for all $P \in \mathcal{P}$
$\mathcal{P}$ can be identified by the family of densities $\{\frac{dP}{d\nu} : P \in \mathcal{P}\}$ or $\{\frac{dP_\theta}{d\nu} : \theta \in \Theta\}$.

Parametric methods: methods designed for parametric models

Example (The k-dimensional normal family).

$$\mathcal{P} = \{N_k(\mu, \Sigma) : \ \mu \in \mathcal{R}^k, \ \Sigma \in \mathcal{M}_k\},$$

where $\mathcal{M}_k$ is a collection of $k \times k$ symmetric positive definite matrices.
This family is dominated by the Lebesgue measure on $\mathcal{R}^k$.
When $k = 1$, $\mathcal{P} = \{N(\mu, \sigma^2) : \ \mu \in \mathcal{R}, \ \sigma^2 > 0\}$.

Nonparametric family: $\mathcal{P}$ is not parametric according to Definition 2.1.
A nonparametric model: the population P is in a given nonparametric family.

Examples of nonparametric family on $(\mathcal{R}^k, \mathcal{B}^k)$:
(1) The joint c.d.f.'s are continuous.
(2) The joint c.d.f.'s have finite moments of order $\leq$ a fixed integer.
(3) The joint c.d.f.'s have p.d.f.'s (e.g., Lebesgue p.d.f.'s).
(4) $k = 1$ and the c.d.f.'s are symmetric.
(5) The family of all probability measures on $(\mathcal{R}^k, \mathcal{B}^k)$.

Nonparametric methods: methods designed for nonparametric models

Semi-parametric models and methods

Statistics and their distributions

Our data set is a realization of a sample (random vector) X from an unknown population P
Statistic $T(X)$: A measurable function T of X; $T(X)$ is a known value whenever X is known.
Statistical analyses are based on various statistics, for various purposes.
X itself is a statistic, but it is a trivial statistic.
The range of a nontrivial statistic $T(X)$ is usually simpler than that of X.
For example, X may be a random n-vector and $T(X)$ may be a random p-vector with a p much smaller than n.
$\sigma(T(X)) \subset \sigma(X)$ and the two σ-fields are the same if and only if T is one-to-one.
Usually $\sigma(T(X))$ simplifies $\sigma(X)$, i.e., a statistic provides a "reduction" of the σ-field.

The "information" within the statistic $T(X)$ concerning the unknown distribution of X is contained in the σ-field $\sigma(T(X))$.
S is any other statistic for which $\sigma(S(X)) = \sigma(T(X))$.
Then, by Lemma 1.2, S is a measurable function of T, and T is a measurable function of S.
Thus, once the value of S (or T) is known, so is the value of T (or S).
It is not the particular values of a statistic that contain the information, but the generated σ-field of the statistic.
Values of a statistic may be important for other reasons.

A statistic $T(X)$ is a random element.
If the distribution of X is unknown, then the distribution of T may also be unknown, although T is a known function.
Finding the form of the distribution of T is one of the major problems in statistical inference and decision theory.

Since T is a transformation of X, tools we learn in Chapter 1 for transformations may be useful in finding the distribution or an approximation to the distribution of $T(X)$.

Example 2.8. Let $X_1, ..., X_n$ be i.i.d. random variables having a common distribution P and $X = (X_1, ..., X_n)$.
The sample mean $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ and sample variance $S^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$ are two commonly used statistics.
Can we find the joint or the marginal distributions of $\bar{X}$ and S^2?
It depends on how much we know about P.

Moments of $\bar{X}$ and S^2
If P has a finite mean μ, then $E\bar{X} = \mu$.
If $P \in \{P_\theta : \theta \in \Theta\}$, then $E\bar{X} = \int x dP_\theta = \mu(\theta)$ for some function $\mu(\cdot)$.
Even if the form of μ is known, $\mu(\theta)$ is still unknown when θ is unknown.
If P has a finite variance σ^2, then $\mathrm{Var}(\bar{X}) = \sigma^2/\mathrm{n}$, which equals $\sigma^2(\theta)/n$ for some function $\sigma^2(\cdot)$ if P is in a parametric family.
With a finite $\sigma^2 = \mathrm{Var}(X_1)$, we can also obtain that $ES^2 = \sigma^2$.
With a finite $E|X_1|^3$, we can obtain $E(\bar{X})^3$ and $\mathrm{Cov}(\bar{X}, S^2)$.
With a finite $E|X_1|^4$, we can obtain $\mathrm{Var}(S^2)$ (exercise).

The distribution of $\bar{X}$
If P is in a parametric family, we can often find the distribution of $\bar{X}$.
See Example 1.20 and some exercises in §1.6.
For example, $\bar{X}$ is $N(\mu, \sigma^2/n)$ if P is $N(\mu, \sigma^2)$;
$n\bar{X}$ has the gamma distribution $\Gamma(n, \theta)$ if P is the exponential distribution $E(0, \theta)$.
If P is not in a parametric family, then it is usually hard to find the exact form of the distribution of $\bar{X}$.
One can use the CLT to obtain an approximation to the distribution of $\bar{X}$.
Applying Corollary 1.2 (for the case of $k = 1$), we obtain that $\sqrt{n}(\bar{X} - \mu) \to_d N(0, \sigma^2)$, where μ and σ^2 are the mean and variance of P, respectively, and are assumed to be finite.
The distribution of $\bar{X}$ can be approximated by $N(\mu, \sigma^2/n)$

The distribution of S^2
If P is $N(\mu, \sigma^2)$, then $(n-1)S^2/\sigma^2$ has the chi-square distribution χ^2_{n-1} (see Example 2.18).
An approximate distribution for S^2 can be obtained from the approximate joint distribution of $\bar{X}$ and S^2 discussed next.

Joint distribution of $\bar{X}$ and S^2
If P is $N(\mu, \sigma^2)$, then $\bar{X}$ and S^2 are independent (Example 2.18).
Hence, the joint distribution of $(\bar{X}, S^2)$ is the product of the marginal distributions of $\bar{X}$ and S^2 given in the previous discussion.
Without the normality assumption, an approximate joint distribution can be obtained.
Assume that $\mu = EX_1$, $\sigma^2 = \mathrm{Var}(X_1)$, and $E|X_1|^4$ are finite.
Let $Y_i = (X_i - \mu, (X_i - \mu)^2)$, $i = 1, ..., n$.

$Y_1, ..., Y_n$ are i.i.d. random 2-vectors with $EY_1 = (0, \sigma^2)$ and variance-covariance matrix

$$\Sigma = \begin{pmatrix} \sigma^2 & E(X_1 - \mu)^3 \\ E(X_1 - \mu)^3 & E(X_1 - \mu)^4 - \sigma^4 \end{pmatrix}.$$

Note that $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i = (\bar{X} - \mu, \tilde{S}^2)$, where $\tilde{S}^2 = n^{-1} \sum_{i=1}^n (X_i - \mu)^2$.
Applying the CLT (Corollary 1.2) to Y_i's, we obtain that

$$\sqrt{n}(\bar{X} - \mu, \tilde{S}^2 - \sigma^2) \to_d N_2(0, \Sigma).$$

Since

$$S^2 = \frac{n}{n-1} \left[\tilde{S}^2 - (\bar{X} - \mu)^2 \right]$$

and $\bar{X} \to_{a.s.} \mu$ (the SLLN), an application of Slutsky's theorem leads to

$$\sqrt{n}(\bar{X} - \mu, S^2 - \sigma^2) \to_d N_2(0, \Sigma).$$

Example 2.9 (Order statistics). Let $X = (X_1, ..., X_n)$ with i.i.d. random components.
Let $X_{(i)}$ be the ith smallest value of $X_1, ..., X_n$.
The statistics $X_{(1)}, ..., X_{(n)}$ are called the *order statistics*.
Order statistics is a set of very useful statistics in addition to the sample mean and variance.
Suppose that X_i has a c.d.f. F having a Lebesgue p.d.f. f.
Then the joint Lebesgue p.d.f. of $X_{(1)}, ..., X_{(n)}$ is

$$g(x_1, x_2, ..., x_n) = \begin{cases} n! f(x_1) f(x_2) \cdots f(x_n) & x_1 < x_2 < \cdots < x_n \\ 0 & \text{otherwise.} \end{cases}$$

The joint Lebesgue p.d.f. of $X_{(i)}$ and $X_{(j)}$, $1 \leq i < j \leq n$, is

$$g_{i,j}(x, y) = \begin{cases} \frac{n![F(x)]^{i-1}[F(y)-F(x)]^{j-i-1}[1-F(y)]^{n-j} f(x) f(y)}{(i-1)!(j-i-1)!(n-j)!} & x < y \\ 0 & \text{otherwise} \end{cases}$$

and the Lebesgue p.d.f. of $X_{(i)}$ is

$$g_i(x) = \frac{n!}{(i-1)!(n-i)!} [F(x)]^{i-1} [1 - F(x)]^{n-i} f(x).$$