Chapter 2: Fundamentals of Statistics
Lecture 10: Models, data, statistics, and sampling distributions

- Data from one or a series of random experiments are collected.
- Planning experiments and collecting data (not discussed here).
- Analysis: extract information from the data and draw conclusions.

- Descriptive data analysis: Summary of the data, such as the mean, median, range, standard deviation, etc., and 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 $(\Omega, \mathscr{F}, P)$, in which P is called the *population*.
- The data set is the realization of a *sample* from P.
- The size of the data set is called the *sample size*.

- A population P is *known* iff $P(A)$ is a known value for every $A \in \mathscr{F}$.
- In a statistical problem, the population P is unknown.
- We deduce properties of P based on the available sample/data.

Read Examples 2.1-2.3

Statistical model

- A *statistical model* is a set of assumptions on the population P and is often postulated to make the analysis possible or easy.
- Postulated models are often based on knowledge of the problem.

Definition 2.1

A set of probability measures P_θ on $(\Omega, \mathscr{F})$ indexed by a *parameter* $\theta \in \Theta$ is said to be a *parametric family* or follow a parametric model iff $\Theta \subset \mathscr{R}^d$ 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*.

$\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is *identifiable* iff $\theta_1 \neq \theta_2$ and $\theta_i \in \Theta$ imply $P_{\theta_1} \neq P_{\theta_2}$, which may be achieved through reparameterization.

Dominated family

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

Example (The k-dimensional normal family)

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

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

Nonparametric family or model

$\mathscr{P}$ is not parametric according to Definition 2.1.

Examples of nonparametric family on $(\mathscr{R}^k, \mathscr{B}^k)$

- All continuous joint c.d.f.'s.
- All joint c.d.f.'s having finite moments of order $\leq$ a fixed integer.
- All joint c.d.f.'s having p.d.f.'s (e.g., Lebesgue p.d.f.'s).
- All symmetric c.d.f.'s.

Definition 2.2 (Exponential families)

A parametric family $\{P_\theta : \theta \in \Theta\}$ dominated by a σ-finite measure ν on $(\Omega, \mathscr{F})$ is called an *exponential family* iff

$$\frac{dP_\theta}{d\nu}(\omega) = \exp\{[\eta(\theta)]^\tau T(\omega) - \xi(\theta)\} h(\omega), \quad \omega \in \Omega,$$

where $\exp\{x\} = e^x$, T is a random p-vector on $(\Omega, \mathscr{F})$ with a fixed positive integer p, η is a function from Θ to $\mathscr{R}^p$, $h \geq 0$ is a Borel function on $(\Omega, \mathscr{F})$, and $\xi(\theta) = \log\left\{\int_\Omega \exp\{[\eta(\theta)]^\tau T(\omega)\} h(\omega) d\nu(\omega)\right\}$.

The representation of an exponential family is not unique.
In an exponential family, consider the parameter $\eta = \eta(\theta)$ and

$$f_\eta(\omega) = \exp\{\eta^\tau T(\omega) - \zeta(\eta)\} h(\omega), \quad \omega \in \Omega, \tag{1}$$

where $\zeta(\eta) = \log\left\{\int_\Omega \exp\{\eta^\tau T(\omega)\} h(\omega) d\nu(\omega)\right\}$.
This is called the *canonical form* for the family, and
$\Xi = \{\eta : \zeta(\eta) \text{ is defined}\}$ is called the *natural parameter space*.
An exponential family in canonical form is a *natural exponential family*.
If $X_1, ..., X_m$ are independent random vectors with p.d.f.'s in exponential families, then the p.d.f. of $(X_1, ..., X_m)$ is again in an exponential family.

If there is an open set contained in the natural parameter space of an exponential family, then the family is said to be of *full rank*.

Example 2.6

The normal family $\{N(\mu,\sigma^2):\ \mu\in\mathscr{R},\sigma>0\}$ is an exponential family, since the Lebesgue p.d.f. of $N(\mu,\sigma^2)$ can be written as

$$\frac{1}{\sqrt{2\pi}}\exp\left\{\frac{\mu}{\sigma^2}x-\frac{1}{2\sigma^2}x^2-\frac{\mu^2}{2\sigma^2}-\log\sigma\right\}.$$

This belongs to an exponential family with $T(x)=(x,-x^2)$, $\eta(\theta)=\left(\frac{\mu}{\sigma^2},\frac{1}{2\sigma^2}\right)$, $\theta=(\mu,\sigma^2)$, $\xi(\theta)=\frac{\mu^2}{2\sigma^2}+\log\sigma$, and $h(x)=1/\sqrt{2\pi}$.

Let $\eta=(\eta_1,\eta_2)=\left(\frac{\mu}{\sigma^2},\frac{1}{2\sigma^2}\right)$.

Then $\Xi=\mathscr{R}\times(0,\infty)$ and we can obtain a natural exponential family of full rank with $\zeta(\eta)=\eta_1^2/(4\eta_2)+\log(1/\sqrt{2\eta_2})$.

A subfamily of the previous normal family, $\{N(\mu,\mu^2):\mu\in\mathscr{R},\mu\neq 0\}$, is also an exponential family with the natural parameter $\eta=\left(\frac{1}{\mu},\frac{1}{2\mu^2}\right)$ and natural parameter space $\Xi=\{(x,y):y=2x^2,\ x\in\mathscr{R},\ y>0\}$.

This exponential family is not of full rank.

Theorem 2.1

Let $\mathscr{P}$ be a natural exponential family with p.d.f. given by (2).

(i) Let $T = (Y, U)$ and $\eta = (\vartheta, \varphi)$, Y and ϑ have the same dimension. Then, Y has the p.d.f. (w.r.t. a σ-finite measure depending on φ)

$$f_\eta(y) = \exp\{\vartheta^\tau y - \zeta(\eta)\}$$

In particular, T has a p.d.f. in a natural exponential family. Furthermore, the conditional distribution of Y given $U = u$ has the p.d.f. (w.r.t. a σ-finite measure depending on u)

$$f_{\vartheta,u}(y) = \exp\{\vartheta^\tau y - \zeta_u(\vartheta)\},$$

which is in a natural exponential family indexed by ϑ.

(ii) If η_0 is an interior point of the natural parameter space, then the m.g.f. of $P_{\eta_0} \circ T^{-1}$ is finite in a neighborhood of 0 and is given by

$$\psi_{\eta_0}(t) = \exp\{\zeta(\eta_0 + t) - \zeta(\eta_0)\}.$$

If f is a Borel function satisfying $\int |f| dP_{\eta_0} < \infty$, then the function $\int f(\omega) \exp\{\eta^\tau T(\omega)\} h(\omega) d\nu(\omega)$ is infinitely often differentiable in a neighborhood of η_0, and the derivatives may be computed by differentiation under the integral sign.

If a $\mathscr{P} = \{f_\theta : \theta \in \Theta\}$ and the set $\{x : f_\theta(x) > 0\}$depends on θ, then $\mathscr{P}$ is not an exponential family.

Definition 2.3 (Location-scale families)

Let P be a known probability measure on $(\mathscr{R}^k, \mathscr{B}^k)$, $\mathscr{V} \subset \mathscr{R}^k$, and $\mathscr{M}_k$ be a collection of $k \times k$ symmetric positive definite matrices.
The family

$$\{P_{(\mu,\Sigma)} : \ \mu \in \mathscr{V}, \ \Sigma \in \mathscr{M}_k\}$$

is called a *location-scale family* (on $\mathscr{R}^k$), where

$$P_{(\mu,\Sigma)}(B) = P\left(\Sigma^{-1/2}(B-\mu)\right), \quad B \in \mathscr{B}^k,$$

$\Sigma^{-1/2}(B-\mu) = \{\Sigma^{-1/2}(x-\mu) : \ x \in B\} \subset \mathscr{R}^k$, and $\Sigma^{-1/2}$ is the inverse of the “square root" matrix $\Sigma^{1/2}$ satisfying $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$.
The parameters μ and $\Sigma^{1/2}$ are called the location and scale parameters, respectively.

$\{P_{(\mu,I_k)} : \ \mu \in \mathscr{R}^k\}$ is a *location family*, where I_k is the identity matrix.
$\{P_{(0,\Sigma)} : \ \Sigma \in \mathscr{M}_k\}$ is a *scale family*.
$\{P_{(\mu,\sigma^2 I_k)} : \ \mu \in \mathscr{R}^k, \sigma > 0\}$ is a location-scale family.

Statistics and their sampling 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.
- A nontrivial statistic $T(X)$ is usually simpler than X.
- Usually $\sigma(T(X)) \subset \sigma(X)$, i.e., $\sigma(T(X))$ simplifies $\sigma(X)$; 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 $\sigma(T(X))$.
- If S is another statistic for which $\sigma(S(X)) = \sigma(T(X))$, then by Lemma 1.2, S and T are functions of each other.
- 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.

Sampling distribution of a statistic

- 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. The sample mean and sample variance

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

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 has a finite variance σ^2, then $\mathrm{Var}(\bar{X}) = \sigma^2/n$ and $ES^2 = \sigma^2$.
- With a finite $E|X_1|^3$, we can obtain $E\bar{X}^3$ and $\mathrm{Cov}(\bar{X}, S^2)$.
- With a finite EX_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 $\bar{X}$. 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 $E(0, \theta)$;
- See Example 1.20 and some exercises in §1.6.

One can use the CLT to get an approximation to the distribution of $\bar{X}$. Applying Corollary 1.2 ($k = 1$), we have $\sqrt{n}(\bar{X} - \mu) \to_d N(0, \sigma^2)$, so that the distribution of $\bar{X}$ can be approximated by $N(\mu, \sigma^2/n)$

Joint distribution of $\bar{X}$ and S^2

If P is $N(\mu, \sigma^2)$, then $\bar{X}$ and S^2 are independent and the joint distribution of $(\bar{X}, S^2)$ can be obtained.

It is enough to show the independence of $\bar{Z}$ and S_Z^2, the sample mean and variance based on $Z_i = (X_i - \mu)/\sigma \sim N(0,1)$, $i = 1,...,n$, because

$$\bar{X} = \sigma \bar{Z} - \mu \quad \text{and} \quad S^2 = \frac{\sigma^2}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})^2 = \sigma^2 S_Z^2$$

Consider the transformation

$$Y_1 = \bar{Z}, \qquad Y_i = Z_i - \bar{Z}, \quad i = 2,...,n,$$

Then

$$Z_1 = Y_1 - (Y_2 + \cdots + Y_n), \quad Z_i = Y_i + Y_1, \quad i = 2,...,n,$$

and

$$\left| \frac{\partial(Z_1,...,Z_n)}{\partial(Y_1,...,Y_n)} \right| = \frac{1}{n}$$

Since the joint pdf of $Z_1,...,Z_n$ is

$$\frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^n z_i^2 \right) \qquad z_i \in \mathscr{R}, i = 1,...,n,$$

the joint pdf of $(Y_1,...,Y_n)$ is

$$\frac{n}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\left(y_1-\sum_{i=2}^n y_i\right)^2\right)\exp\left(-\frac{1}{2}\sum_{i=2}^n (y_i+y_1)^2\right)$$

$$= \frac{n}{(2\pi)^{n/2}}\exp\left(-\frac{n}{2}y_1^2\right)\exp\left(-\frac{1}{2}\left[\sum_{i=2}^n y_i^2+\left(\sum_{i=2}^n y_i\right)^2\right]\right) \qquad \begin{array}{c} y_i\in\mathscr{R} \\ i=1,...,n. \end{array}$$

Since the first exp factor involves y_1 only and the second exp factor involves $y_2,...,y_n$, we conclude that Y_1 is independent of $(Y_2,...,Y_n)$. Since

$$Z_1-\bar{Z}=-\sum_{i=2}^n (Z_i-\bar{Z})=-\sum_{i=2}^n Y_i \quad \text{and} \quad Z_i-\bar{Z}=Y_i, \quad i=2,...,n,$$

we have

$$S_Z^2=\frac{1}{n-1}\sum_{i=1}^n (Z_i-\bar{Z})^2=\frac{1}{n-1}\left(\sum_{i=2}^n Y_i\right)^2+\frac{1}{n-1}\sum_{i=2}^n Y_i^2$$

which is a function of $(Y_2,...,Y_n)$.

Hence, $\bar{Z}$ and S_Z^2 are independent.

Note that

$$(n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu + \mu - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 + n(\mu - \bar{X})^2$$

Then

$$n\left(\frac{\bar{X}-\mu}{\sigma}\right)^2 + \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n} Z_i^2 \qquad (2)$$

Since $Z_i \sim N(0,1)$ and $Z_1, ..., Z_n$ are independent, $\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$
Since $\sqrt{n}(\bar{X}-\mu)/\sigma \sim N(0,1)$, $n[(\bar{X}-\mu)/\sigma]^2 \sim \chi_1^2$.
The left hand side of (2) is a sum of two independent random variables and, hence, if $f(t)$ is the mgf of $(n-1)S^2/\sigma^2$, then the mgf of the sum on the left hand side of (2) is $(1-2t)^{-1/2}f(t)$
Since the right hand side of (2) has mgf $(1-2t)^{-n/2}$, we must have

$$f(t) = (1-2t)^{-n/2}/(1-2t)^{-1/2} = (1-2t)^{-(n-1)/2} \qquad t < 1/2$$

This is the mgf of χ_{n-1}^2, hence $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$.

Joint distribution of $\bar{X}$ and S^2

If P is $N(\mu, \sigma^2)$, then $\bar{X}$ and S^2 are independent, $\bar{X} \sim N(\mu, \sigma^2/n)$ and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

Without the normality assumption, we consider an approximation.

Assume that $\mu = EX_1$, $\sigma^2 = \mathrm{var}(X_1)$, and $E|X_1|^4$ are finite.
If $Y_i = (X_i - \mu, (X_i - \mu)^2)$, $i = 1, ..., n$, then $Y_1, ..., Y_n$ 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 $X_{(j)}$, $1 \le i < j \le n$, 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

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

Example.

Let $X_1, ..., X_n$ be a random sample from *uniform*$(0,1)$.
We want to find the distribution of $X_1/X_{(1)}$.
For $s > 1$,

$$
\begin{aligned}
P\left(\frac{X_1}{X_{(1)}} > s\right) &= \sum_{i=1}^{n} P\left(\frac{X_1}{X_{(1)}} > s, X_{(1)} = X_i\right) \\
&= \sum_{i=2}^{n} P\left(\frac{X_1}{X_{(1)}} > s, X_{(1)} = X_i\right) \\
&= (n-1)P\left(\frac{X_1}{X_{(1)}} > s, X_{(1)} = X_n\right) \\
&= (n-1)P(X_1 > sX_n, X_2 > X_n, ..., X_{n-1} > X_n) \\
&= (n-1)P(sX_n < 1, X_1 > sX_n, X_2 > X_n, ..., X_{n-1} > X_n) \\
&= (n-1)\int_0^{1/s}\left[\int_{sx_n}^{1}\left(\prod_{i=2}^{n-1}\int_{x_n}^{1} dx_i\right) dx_1\right] dx_n \\
&= (n-1)\int_0^{1/s}(1-x_n)^{n-2}(1-sx_n)dx_n
\end{aligned}
$$