Chapter 6 Introduction to Statistics

6.1 Population and Random Sample

- **Definition: Random Sample.** A random sample, denoted as $\mathbf{X}^n = (X_1, \cdots, X_n)$, is a sequence of n random variables $X_1, \cdots, X_n$. A realization of the random sample $\mathbf{X}^n$, denoted as $\mathbf{x}^n = (x_1, \cdots, x_n)$, is called a data set generated from $\mathbf{X}^n$ or a sample point of $\mathbf{X}^n$. The positive integer n is called the sample size. A random sample $\mathbf{X}^n$ constitutes the sample space of the random sample $\mathbf{X}^n$.

- **Remark:** The joint pdf or PMF of random vector $(X_1, \cdots, X_n)$ is

$$f_{\mathbf{X}^n}(\mathbf{x}^n) = \prod_{i=1}^{n} f_{X_i|\mathbf{X}^{i-1}}(x_i|\mathbf{x}^{i-1})$$

$$f(x_1, \cdots, x_n) = \prod_{i=1}^{n} f(x_i). IID$$

6.1 Population and Random Sample

- **Definition: IID Random Sample.** The sequence of random variables $X_1, \cdots, X_n$ is called an independent and identically distributed (IID) random sample of size n from the population $F_X(x)$ if:

(1) $X_1, \cdots, X_n$ are mutually *independent* random variables, and

(2) each X_i has the same marginal distribution $F_X(x)$.

- **Question:**
  - What is the interpretation and implication of an IID random sample?
  - How to define the population if the random variables $X_1, \cdots, X_n$ in the sample are not identically distributed?
  - How to extract information from a data set $\mathbf{x}^n$?

6.1 Population and Random Sample

- **Example: Tossing Coins** Suppose we throw n coins. Let X_i denote the outcome of throwing the $i-$th coin, with $X_i = 1$ for head, and $X_i = 0$ for tail. Then $\mathbf{X}^n = X_1, ..., X_n{}'$ constitutes a random sample. We will obtain a sequence of real numbers, such as $\mathbf{x}^n = (1, 1, 0, 0, 1, 0..., 1)$.

- **Example: GDP Annual Growth Rate.** Let X_i denote the Chinese GDP growth rate in year i, from 1953 to 2010. Then $\mathbf{X}^n = X_1, ..., X_n{}'$ constitutes a random sample with sample size $n = 58$.

6.1 Population and Random Sample

- **Definition: Statistic.** Let $\mathbf{X}^n = (X_1, \cdots, X_n)$ be a random sample of size n from a population. A statistic $T(\mathbf{X}^n) = T(X_1, \cdots, X_n)$ is a real-valued or vector valued function of a random sample $\mathbf{X}^n$. Then the random variable or random vector $Y = T(X_1, \cdots, X_n)$ is called a *statistic*.

- **Examples:**
  - Sample mean: $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.
  - Sample variance: $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.
  - Sample standard deviation: $S_n = \sqrt{S_n^2}$.

- **Definition: Sampling Distribution.** The probability distribution of a statistic $T(\mathbf{X^n})$ is called the *sampling distribution* of $T(\mathbf{X}^n)$.

- **Definition: Log-likelihood Function.** Let $X_1, \cdots, X_n$ be an IID random sample of size n from the population $f(x;\theta)$. Then the logarithm of the joint pmf/pdf of $(X_1, \cdots, X_n)$

$$L(\theta \mid X_1, \cdots, X_n) = \log \prod_{i=1}^{n} f(X_i;\theta) = \sum_{i=1}^{n} \log f(X_i;\theta)$$

is called the *log-likelihood function* of θ; conditional on the random sample $X_1, \cdots, X_n$.

- **Remark:** $L(\theta \mid X_1, \cdots, X_n)$ is not a statistic, because it depends on the parameter θ.

6.2 The Sampling Distribution of the Sample Mean

- **Definition: Sample Mean.** Suppose $\mathbf{X}^n = (X_1, \cdots, X_n)$ is a random sample from a population with mean μ and variance σ^2. Then

$$T(\mathbf{X}^n) \equiv \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is the sample mean for the random sample $\mathbf{X}^n$.

- **Theorem.** Suppose $\mathbf{X}^n$ is a random sample. Then

$$\bar{X}_n = \arg \min_{-\infty<a<\infty} \sum_{i=1}^{n}(X_i - a)^2.$$

6.2 The Sampling Distribution of the Sample Mean

- **Another Version.** Let $x_1, \cdots, x_n$ be any number and $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$, then

$$\min_a \sum_{i=1}^n (x_i - a)^2 = \sum_{i=1}^n (x_i - \overline{x}_n)^2.$$

- **Remarks:**
  - $\sum_{i=1}^n (x_i - \overline{x}_n)^2 = \sum_{i=1}^n x_i^2 - n\overline{x}_n^2$.
  - Define $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x}_n)^2$.
  - $\widehat{a} = \overline{x}_n$ is the *ordinary least square* (OLS) estimator for the linear regression model

$$x_i = a + \varepsilon_i,$$

where $\{\varepsilon_i\}$ is an i.i.d. sequence with $E(\varepsilon_i) = 0$.

6.2 The Sampling Distribution of the Sample Mean

- **Lemma:** Let $X_1, \cdots, X_n$ be an IID random sample of a population and let $g(x)$ be a function such that $\text{Var}(X)$ exist. Then

$$E\left[\sum_{i=1}^{n} g(X_i)\right] = nE[g(X_1)],$$

and

$$\text{Var}\left[\sum_{i=1}^{n} g(X_i)\right] = n\text{Var}[g(X_1)].$$

6.2 The Sampling Distribution of the Sample Mean

- **Theorem:** Suppose $X_1, \cdots, X_n$ are identically distributed random variables with the same mean μ. Then

$$E(\overline{X}_n) = \mu.$$

- **Theorem:** Suppose $X_1, \cdots, X_n$ are IID random variables with $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then for all $n \geq 1$,

$$\mathrm{Var}(\overline{X}_n) = \sigma^2/n.$$

- **Remark:** The sample mean and the sample variance can be considered as the approximations of the true mean and the true variance of the population $f(x)$, respectively.

6.2 The Sampling Distribution of the Sample Mean

- **Theorem.** Let $X_1, \cdots, X_n$ be an IID random sample from a population with mgf $M_X(t)$. Then the mgf of the sample mean is

$$M_{\overline{X}_n}(t) = [M_X(t/n)]^n.$$

- **Theorem.** Suppose $X_1, \cdots, X_n$ are IID normally distributed with mean μ and variance $\sigma^2 < \infty$. Define the standardized sample mean

$$Z_n = \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{var(\bar{X}_n)}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}.$$

Then $Z_n \sim N(0,1)$ for all $n \geq 1$.

- **Example:** Let $X_1, \cdots, X_n$ be an IID random sample from $N(\mu, \sigma^2)$, then the mgf of $\overline{X}_n$ is

$$M_{\overline{X}_n}(t) = exp\{\mu t + \frac{(\sigma^2/n)t^2}{2}\},$$

and $\overline{X}_n$ has a $N(\mu, \sigma^2/n)$ distribution.

6.2 The Sampling Distribution of the Sample Mean

- **Example:** Let $X_1, \cdots, X_n$ be an IID random sample from Gamma(α, β), then the mgf of $\overline{X}_n$ is

$$M_{\overline{X}_n}(t) = (1 - \beta\frac{t}{n})^{-n\alpha},$$

and $\overline{X}_n$ has a Gamma$(n\alpha, \beta/n)$ distribution.

- **Example:** Let $Z_1, \cdots, Z_n$ be an IID random sample from Cauchy$(0, \sigma)$ distribution with the pdf $f_X(x) = \frac{1}{\pi\sigma}\frac{1}{1+(x/\sigma)^2}$, then $\overline{Z}_n$ also has a Cauchy$(0, \sigma)$ distribution.

- **Proof.**

  – If $U \sim$ Cauchy$(0, \eta)$, $V \sim$ Cauchy$(0, \tau)$, then $U + V \sim$ Cauchy$(0, \eta + \tau)$ (Exercise 5.7, Casella & Berger).

  – Because $Z_1 + \cdots + Z_n \sim$ Cauchy$(0, n\sigma)$, so $\overline{Z}_n \sim$ Cauchy$(0, \sigma)$.

6.3 The Sampling Distribution of the Sample Variance

- **Sample Variance Estimator.** $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$

- **Questions:**
  - What is the mean of S_n^2?
  - What is the variance of S_n^2?
  - What is the sampling distribution of S_n^2?

- **Theorem.** Suppose $\mathbf{X}^n = (X_1, \cdots, X_n)$ is an IID random sample from a population with (μ, σ^2). Then

$$E(S_n^2) = \sigma^2.$$

6.3 The Sampling Distribution of the Sample Variance

- **Lemma.** Let $Z_1, \cdots, Z_k$ be IID N(0,1) random variables. Then

$$\sum_{i=1}^{k} Z_i^2 \sim \chi_k^2.$$

That is, the sum of squares of k independent N(0,1) random variable follows a χ_k^2 distribution.

- **Theorem.** Suppose $\mathbf{X}^n = (X_1, \cdots, X_n)$ is an IID $N(\mu, \sigma^2)$ random sample. Then for each $n > 1$,

$$\frac{(n-1)S_n^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma^2} \sim \chi_{n-1}^2,$$

where χ_{n-1}^2 is a chi-square distribution with $n-1$ degrees of freedom.

6.3 The Sampling Distribution of the Sample Variance

- **Lemma.** Let $X_1, \cdots, X_n$ be an IID random sample with $X_i \sim N(\mu, \sigma^2)$. For constants a_{ij} and b_{rj}, $j = 1, \cdots, n$,$i = 1, ..., \nu$, $r = 1, \cdots, m$, define

$$\begin{aligned} U_i &= \sum_{j=1}^{n} a_{ij} X_i, \qquad i = 1, \cdots, \nu, \\ V_r &= \sum_{j=1}^{n} b_{rj} X_i, \qquad r = 1, \cdots, m. \end{aligned}$$

where $\nu + m \leq n$. Then

(1) The random variables U_i and V_r are independent iff $\mathrm{Cov}(U_j, V_r) = 0$.

(2) The random vectors $(U_1, \cdots, U_\nu)$,and $(V_1, \cdots, V_m)$ are independent are independent if and only if U_i is independent of V_r for all $i = 1, ..., \nu$, $r = 1, ..., m$.

6.3 The Sampling Distribution of the Sample Variance

- **Definition:** Random vectors $(U_1, \cdots, U_k)$ and $(V_1, \cdots, V_m)$ are *independent* if for any Borel sets $A \subset \mathbb{R}^k$ and $B \subset \mathbb{R}^m$,

$$\begin{aligned} & P[(U_1, \cdots, U_k) \in A, (V_1, \cdots, V_m) \in B] \\ & \qquad = P[(U_1, \cdots, U_k) \in A] P[(V_1, \cdots, V_m) \in B]. \end{aligned}$$

- **Remarks:**
  - $(U_1, \cdots, U_k)$ and $(V_1, \cdots, V_m)$ are independent if and only if the joint pdf/pmf can be written as

$$f_{UV}(u_1, \cdots, u_k, v_1, \cdots, v_m) = h(u_1, \cdots, u_k) g(v_1, \cdots, v_m)$$

for all $(u_1, \cdots, u_k) \in \mathbb{R}^k$ and $(v_1, \cdots, v_m) \in \mathbb{R}^m$.

6.3 The Sampling Distribution of the Sample Variance

- **Remarks:**
  - If $(U_1, \cdots, U_k)$ and $(V_1, \cdots, V_m)$ are independent, then for any $g(\cdot)$ and $h(\cdot)$, $g(U_1, \cdots, U_k)$ and $h(V_1, \cdots, V_m)$ are independent random variables.
  - If $(U_1, \cdots, U_k)$ and $(V_1, \cdots, V_m)$ are independent, then every pair of U_i and V_r are independent.
  - The converse may not be true. For example, in cases X_1, X_2, X_3 are pairwise independent, but not jointly independent, let $U = (X_1, X_2)$, $V = X_3$.

6.3 The Sampling Distribution of the Sample Variance

- **Lemma.** Let $X_1, \cdots, X_n$ be independent random variables, $X_i \sim N(\mu_i, \sigma_i^2)$. For constants a_{ji} and b_{ri}, $j = 1, \cdots, k$, $r = 1, \cdots, m$, where $k + m \leq n$, define

$$\begin{aligned} U_j &= \sum_{i=1}^{n} a_{ji} X_i, \qquad j = 1, \cdots, k, \\ V_r &= \sum_{i=1}^{n} b_{ri} X_i, \qquad r = 1, \cdots, m. \end{aligned}$$

Then

(1) $(U_1, \cdots, U_k, V_1, \cdots, V_m)$ are jointly normal distributed.

(2) U_j and V_r are independent iff $\mathrm{Cov}(U_j, V_r) = \sum_{i=1}^{n} a_{ji} b_{ri} \sigma_i^2 = 0$.

(3) $(U_1, \cdots, U_k)$ and $(V_1, \cdots, V_m)$ are independent if and only if U_j is independent of V_r for all j, r.

6.3 The Sampling Distribution of the Sample Variance

- **Theorem.** Suppose $\mathbf{X}^n$ is an IID $N(\mu, \sigma^2)$ random sample. Then for any $n > 1$, S_n^2 and $\bar{X}_n$ are mutually independent.

- **Remark.** The sum of squares $\sum_{i=1}^n (X_i - \bar{X}_n)^2$ has only $n-1$ degrees of freedom.

- **Ordinary Least Square(OLS).** In classical linear regression model $Y_i = X_i'\beta + \varepsilon_i$, where X_i is a $p \times 1$ explanatory vector, β is a $p \times 1$ parameter vector, $\{\varepsilon_i\}_{i=1}^n$ is an IID sequence from $N(0, \sigma_\varepsilon^2)$. For simplicity, assume that $\{X_i\}_{i=1}^n$ are nonstochastic.

- What is the distribution of each Y_i?

- Find the OLS estimator $\hat{\beta}$ for β.

- Find the distribution of $\hat{\beta}$ under some regularity conditions

- Find the residual variance estimator for σ_ε^2.
- How about if X_is are random variables?

Hint:

- $Y_i \sim N(X_i'\beta, \sigma_\varepsilon^2)$.
- Then OLS estimator for β is

$$\hat{\beta} = \big(\sum_{i=1}^{n} X_i X_i'\big)^{-1} \sum_{i=1}^{n} X_i Y_i = (X'X)^{-1} X'Y,$$

  by $\hat{\beta} = \arg\min_\beta \sum_{i=1}^n (Y_i - X_i'\beta)^2$.

- Under a set of regularity conditions, $\hat{\beta} - \beta_0 \sim N(0, \sigma_\varepsilon^2 (X'X)^{-1})$.
- We can use $S_\varepsilon^2 = \frac{1}{n-p} \sum_{i=1}^n (Y_i - X_i'\hat{\beta})^2$ to estimate σ_ε^2.
- If X_is are random variables, we should consider $Y|X_i$, conditional case.

6.3 The Sampling Distribution of the Sample Variance

- **Definition: Chi Squared Distribution.** A random variable follows *chi squared* distribution with *degrees of freedom* p, denoted by χ_p^2, if its pdf has the form

$$f(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}, \quad 0 < x < \infty.$$

- **Facts about chi squared random variables.**

  (a) If X is a $N(0,1)$ random variable, then $X^2 \sim \chi_1^2$.

  (b) If $Z_1, \cdots, Z_n$ are independent and $Z_i \sim \chi_{p_i}^2$, then $Z_1 + \cdots + Z_n \sim \chi^2_{p_1+\cdots+p_n}$.

- **Proof.** (a) Calculate the pdf of $Y = X^2$. (b) Z_i follows Gamma$(p_i/2, 2)$ with mgf $(1-2t)^{-p_i/2}$, then the mgf of $Z_1 + \cdots + Z_n$ is $(1-2t)^{-(p_1+\cdots+p_n)/2}$. $Z_1 + \cdots + Z_n$ follows Gamma$((p_1 + \cdots + p_n)/2, 2)$.

- **Theorem.** Suppose $\mathbf{X}^n = (X_1, \cdots, X_n)$ is an IID $N(\mu, \sigma^2)$ random sample. Then for all $n > 1$,
$$\text{var}(S_n^2) = \frac{2\sigma^4}{n-1}.$$
The fact that $\text{var}(S_n^2) = 2\sigma^4/(n-1)$ and $E(S_n^2) = \sigma^2$ implies that
$$\text{MSE}(S_n^2) = E(S_n^2 - \sigma^2)^2 = \text{var}(S_n^2) = \frac{2\sigma^4}{n-1} \to 0 \text{ as } n \to \infty.$$

6.4 Student's t-Distribution

- **Definition:** If $U \sim N(0,1)$, $V \sim \chi^2_\nu$, U and V are independent, then $T = U/\sqrt{V/\nu}$ has *Student's t distribution* with ν degrees of freedom, and we write $T \sim t_\nu$. It has pdf

$$f_T(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{(\nu\pi)^{1/2}} \frac{1}{(1+t^2/\nu)^{(\nu+1)/2}}, \quad -\infty < t < \infty.$$

- **Remarks:**
  - t_1 has no mean, t_2 has no variance.
  - If $Y \sim t_\nu$, then $E(Y) = 0$, $\mathrm{Var}(Y) = \nu/(\nu-2)$.
  - The density can be obtained by $T = U/\sqrt{V/\nu}$, $R = U$,and then integrating our R.
  - $\lim_{\nu\to\infty} f_T(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$.

6.4 Student's t-Distribution

- **Lemma: Properties of the Student t_ν Distribution.**

  (1) The PDF of t_ν is symmetric about 0.

  (2) t_ν has a heavier distributional tail than $N(0,1)$.

  (3) Only the first $\nu - 1$ moments exist. The mean $\mu = 0$, and the variance $\sigma^2 = \nu/(\nu-2)$ when $\nu > 2$. The MGF does not exist for any given k.

  (4) When $\nu = 1$, $t_1 \sim$ Cauchy(0,1).

  (5) $t_\nu \to N(0,1)$ as $\nu \to \infty$.

- **Theorem:** Let $\mathbf{X}^n = (X_1, \cdots, X_n)$ be an IID random sample from $N(\mu, \sigma^2)$, then for all $n > 1$, the standardized sample mean

$$\frac{\bar{X} - \mu}{S_n/\sqrt{n}} = \frac{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S_n^2}{\sigma^2}/(n-1)}} \sim \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}} \sim t_{n-1},$$

  where t_{n-1} is the Student t distribution with $n-1$ degrees of freedom.

- **Example: Confidence Interval Estimation for Population Mean μ.** Suppose there is an IID $N(\mu, \sigma^2)$ random sample $\mathbf{X}^n = (X_1, \cdots, X_n)$ of size n, where both μ and σ^2 are unknown. We are interested in constructing a confidence interval estimator for μ at the $(1-\alpha)\%$ confidence level.
  - $(1-\alpha)\%$ confidence level for μ is defined an random interval $[\hat{L}_n, \hat{U}_n]$, such that $P(\hat{L}_n < \mu < \hat{U}_n) = 1-\alpha$.
  - $P(|\frac{\sqrt{n}(\bar{X}_n-\mu)}{S_n}| \leq C_{n-1,\alpha/2}) = 1-\alpha$.
  - $P(\bar{X}_n - \frac{S_n}{\sqrt{n}} C_{t_{n-1},\alpha/2} < \mu < \bar{X}_n + \frac{S_n}{\sqrt{n}} C_{t_{n-1},\alpha/2}) = 1-\alpha$.
  - The $(1-\alpha)\%$ confidence interval for μ is $[\bar{X}_n - \frac{S_n}{\sqrt{n}} C_{t_{n-1},\alpha/2}, \bar{X}_n + \frac{S_n}{\sqrt{n}} C_{t_{n-1},\alpha/2}]$

- **Example: Hypothesis Testing on Population Mean: The t-test.** Suppose there is an IID $N(\mu, \sigma^2)$ random sample $\mathbf{X}^n = (X_1, \cdots, X_n)$ of size n, and we are interested in testing the hypothesis

$$\mathcal{H}_0 : \mu = \mu_0,$$

where μ is a given(known) constant (e.g., $\mu_0 = 0$). How can we test this hypothesis?

  - Statistic: $\bar{X}_n - \mu_0 = (\bar{X}_n - \mu) + (\mu - \mu_0)$.
  - How far away $\bar{X}_n - \mu_0$ is from zero will be considered as "sufficiently large" in absolute value ?
  - Feasible statistic $T(\mathbf{X}^n) = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$.
  - Decision rule: Rejection region, p-value.

6.5 Snedecor's F Distribution

- **Definition: The F Distribution.** Let $U \sim \chi^2_p$, $V \sim \chi^2_q$, U and V are independent, then $F = (U/p)/(V/q)$ has the F *distribution* with p and q degrees of freedom, denoted by $F_{p,q}$, with pdf

$$f_F(x) = \frac{\Gamma(\frac{p+q}{2})}{\Gamma(\frac{p}{2})\Gamma(\frac{q}{2})} \left(\frac{p}{q}\right)^{p/2} \frac{x^{p/2-1}}{[1+(p/q)x]^{(p+q)/2}}, \quad 0 < x < \infty.$$

PDF can be obtained by $F = (U/p)/(V/q)$, $G = U$, and integrating out G.

- **Lemma:**
  - If $X \sim F_{p,q}$, then $1/X \sim F_{q,p}$.
  - If $X \sim t_q$, then $X^2 \sim F_{1,q}$.
  - If $q \to \infty$, then $p \cdot F_{p,q} \to \chi^2_p$.

6.5 Snedecor's F Distribution

- **Theorem:** Let $X_1, \cdots, X_n$ be a random sample from a $N(\mu_X, \sigma_X^2)$ population, and let $Y_1, \cdots, Y_m$ from an independent $N(\mu_Y, \sigma_Y^2)$ population. The random variable $F = (S_{n,X}^2/\sigma_X^2)/(S_{m,Y}^2/\sigma_Y^2) \sim F_{n-1,m-1}$.

- **Example: Hypothesis Testing on equality of Population Variances.** Let $\mathbf{X}^n = (X_1, \cdots, X_n)$ be an IID random sample of size n from a $N(\mu_X, \sigma_X^2)$ population, and $\mathbf{Y}^m = (Y_1, \cdots, Y_m)$ be an IID random sample of size m from a $N(\mu_Y, \sigma_Y^2)$ population. Assume that $\mathbf{X}^n$ and $\mathbf{Y}^m$ are independent. Suppose we are interested in comparing variability of the population, i.e. interested in testing whether $\mathcal{H}_0 : \sigma_X^2 = \sigma_Y^2$ holds. Test statistic is $(S_{n,X}^2/S_{m,Y}^2)$

6.5 Snedecor's F Distribution

F-Test. Consider the classical linear regression model

$$Y_i = X_i'\beta + Z_i'\gamma + \varepsilon_i,$$

where β is $p \times 1$ parameter vector, and γ is $q \times 1$ parameter vector, $\{\varepsilon_i\}$ is a sequence of IID $N(0, \sigma_\varepsilon^2)$ random variables and is independent of $\mathbf{X} = (X_1, ..., X_n)'$ and $\mathbf{Z} = (Z_1, ..., Z_n)'$. We are interested in

$$H_0 : \gamma = 0.$$

- Under H_0, $Y_i = X_i'\beta + \varepsilon_i$,, $\tilde{\beta} = \arg\min_\beta \sum_{i=1}^n (Y_i - X_i'\beta)^2$, $S_R^2 = \frac{1}{n-p} \sum_{i=1}^n (Y_i - X_i'\tilde{\beta})^2$.
- Under H_1, $Y_i = X_i'\beta + Z_i'\gamma + \varepsilon_i$, $(\hat{\beta}, \hat{\gamma}) = \arg\min_{\beta,\gamma} \sum_{i=1}^n (Y_i - X_i'\beta - Z_i'\gamma)^2$, $S_U^2 = \frac{1}{n-p-1} \sum_{i=1}^n (Y_i - X_i'\hat{\beta} - Z_i'\hat{\gamma})^2$.
- Under H_0, Test statistic $F = \frac{[(n-p)S_R^2 - (n-p-q)S_U^2]/q}{(n-p-q)S_U^2/(n-p-q)} \sim F_{q,n-p-q}$

6.6 Sufficient Statistics

- **Sufficiency Principle.** A statistic $T(X_1, \cdots, X_n)$ is a *sufficient statistic* for a parameter θ if any inference about θ **only** depends on the value of $T(\boldsymbol{X})$. That is, if $\boldsymbol{x} = (x_1, \cdots, x_n)$ and $\boldsymbol{y} = (y_1, \cdots, y_n)$ 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.

- **Interpretation:** When a sample $X_1, \cdots, X_n$ is given, there might be too much irrelevant information if we are only interested in knowing the value of some parameter θ. The sufficiency principle provides a method of data reduction that "throw away" the irrelevant information and maintain only the essential information $T(\boldsymbol{X})$.

6.6 Sufficient Statistics

- **Definition: Sufficient Statistic.** Let $\boldsymbol{X}^n = (X_1, \cdots, X_n)$ be a random sample from some population $f(x;\theta)$ with the parameter θ. A statistics $T(\boldsymbol{X}^n)$ is a *sufficient statistic* for θ if the conditional distribution of the sample $\boldsymbol{X}^n = \boldsymbol{x}^n$ given the value of $T(\boldsymbol{X}^n) = T(\boldsymbol{x}^n)$ does not depend on θ, *i.e.*,

$$f_{\boldsymbol{X}^n|T(\boldsymbol{X}^n)}[\boldsymbol{x}^n|T(\boldsymbol{x}^n),\theta)] = h(\boldsymbol{x}^n) \text{ for all possible } \theta,$$

where the left hand side is the conditional PMF/PDF of $\boldsymbol{X}^n = \boldsymbol{x}^n$ given $T(\boldsymbol{X}^n) = T(\boldsymbol{x}^n)$, which generally depends on θ. The right hand side $h(\boldsymbol{x}^n)$ does not depend on θ; it is a function $\boldsymbol{x}^n$ only.

- **Example:** The *maximum likelihood estimate* (MLE) of θ is obtained by maximizing the *likelihood* function $L(\theta) = f_X(\boldsymbol{x};\theta)$ as a function of θ given observation $\boldsymbol{x}$. If $T(\boldsymbol{X})$ is a sufficient statistic, then $L(\theta) = f_X(\boldsymbol{x};\theta)$ is maximized at the same point at which $f_T[T(\boldsymbol{x});\theta]$ is maximized.

6.6 Sufficient Statistics

- **Theorem: Factorization Theorem.** Let $f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta)$ be the pmf or pdf of $\boldsymbol{X}^n$ with parameter θ. Then $T(\boldsymbol{X}^n)$ is a sufficient statistics for θ if and only if there exist functions $g(t, \theta)$ and $h(\boldsymbol{x}^n)$ such that for any sample points $\boldsymbol{x}^n$ in the sample space of $\boldsymbol{X}^n$ and for any parameter points $\theta \in \Theta$,

$$f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta) = g[T(\boldsymbol{x}^n), \theta]h(\boldsymbol{x}^n),$$

  where $g(t, \theta)$ depends on parameter θ but $h(\boldsymbol{x}^n)$ does not depend on parameter θ.

- **Theorem: Invariance Principle.** Suppose $T(\boldsymbol{X}^n)$ is a sufficient statistic for θ, then any 1-1 function $G(\boldsymbol{X}^n) = r[T(\boldsymbol{X}^n)]$ is also a sufficient statistic for θ, and a sufficient statistic for the transformed parameter $r(\theta)$.

- **Example:** It is always true that the complete random sample $(X_1, \cdots, X_n)$ is a sufficient statistic for parameter θ.

- **Example:** Let $X_1, \cdots, X_n$ be i.i.d. Bernoulli random variables with parameter θ. The pmf of X_i is $f(x, \theta) = \theta^x(1-\theta)^{1-x}$ for $x = 0, 1$. We have

$$\begin{aligned} f(x_1, \cdots, x_n, \theta) &= \prod_{i=1}^{n} f(x_i, \theta) \\ &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} h(\boldsymbol{x}), \end{aligned}$$

where $h(\boldsymbol{x}) = 1$ if all $x_i = 0$ or 1, otherwise $h(\boldsymbol{x}) = 0$. Therefore, $T(\boldsymbol{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic for θ.

- **Example: Normal Sufficient Statistic.** Let $X_1, \cdots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. We have

  –
  $$\begin{aligned} f(x_1, \cdots, x_n, \theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \overline{x}_n + \overline{x}_n - \mu\right)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} exp\left\{-\frac{1}{2\sigma^2}\left[n\left(\overline{x}_n - \mu\right)^2 + \sum_{i=1}^{n}\left(x_i - \overline{x}_n\right)^2\right]\right\}. \end{aligned}$$

  – $\overline{X}_n$ is a sufficient statistic for μ if σ^2 is known.

  – When both μ and σ^2 are unknown parameters, $T = (\overline{X}_n, S_n^2)$ is a sufficient statistic for (μ, σ^2).

- **Example: Uniform Sufficient Statistic.** Let $X_1, \cdots, X_n$ be a random sample from the discrete uniform distribution on $1, \cdots, \theta$, where θ is a positive integer.

  - The joint pmf of $X_1, \cdots, X_n$ is

$$f(x_1, \cdots, x_n, \theta) = \begin{cases} \frac{1}{\theta^n} & x_i \in \{1, \cdots, \theta\}, \\ 0 & \text{otherwise.} \end{cases}$$

  - The joint pmf can be rewritten as

$$f(x_1, \cdots, x_n, \theta) = \frac{1}{\theta^n} I(x_{(n)} \leq \theta) h(x_1, \cdots, x_n),$$

    where $h(\boldsymbol{x}) = 1$ if all x_i are positive integers, otherwise $h(\boldsymbol{x}) = 0$.

  - $X_{(n)}$ is a sufficient statistic for θ.

- **Definition: Exponential Family.** A family of pdf or pmf is called an *exponential family* if it can be expressed as

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

- **Examples:**
  - Discrete exponential families: Bernoulli, binomial, Poisson.
  - Continuous exponential families: normal, gamma, beta.
  - The expression needs to hold for all $x \in \mathbb{R}$. In cases that the support $A = \{x : f(x, \theta) > 0\}$ is not $\mathbb{R}$, we can incorporate the indicator function $I_A(x)$ in $h(x)$.
  - In general, the support A of an exponential family can not depend on θ.

6.6 Sufficient Statistics

- **Theorem:** Let $\boldsymbol{X}^n = (X_1, \cdots, X_n)$ be an IID random sample from a pdf or a pmf that belongs to an exponential family

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

then

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

is a sufficient statistic for θ.

- **Problem:** In any model, there are numerous sufficient statistics for θ. What is the most efficient sufficient statistic that achieves the most data reduction while still containing all the information about θ?

- **Definition: Minimal Sufficient Statistic.** A sufficient statistic $T(\boldsymbol{X}^n)$ is called a *minimal sufficient statistic* for parameter θ if, for any other sufficient statistic $G(\boldsymbol{X}^n)$, $T(\boldsymbol{X}^n)$ is a function of $G(\boldsymbol{X}^n)$. That is, for any sufficient statistic $G(\boldsymbol{X}^n)$, there always exists some function $r(\cdot)$ such that $T(\boldsymbol{X}^n) = r[G(\boldsymbol{X}^n)]$.

- **Remarks:**
  - If for any two sample points $\boldsymbol{x}$ and $\boldsymbol{y}$, $T'(\boldsymbol{x}) = T'(\boldsymbol{y})$ implies $T(\boldsymbol{x}) = T(\boldsymbol{y})$, then $T(\boldsymbol{X})$ is a function of $T'(\boldsymbol{X})$.
  - Minimal sufficient statistic may not exist.
  - If $T(\boldsymbol{X})$ is a minimal sufficient statistic, then for any 1-1 function r, $T^*(\boldsymbol{x}) = r(T(\boldsymbol{x}))$ is also a minimal sufficient statistic.

- **Theorem.** Let $f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta)$ be the PMF/PDF of a random sample $\boldsymbol{X}^n$. Suppose there exists a function $T(\boldsymbol{X}^n)$ such that, for two sample points x^n and y^n in the sample space of $\boldsymbol{X}^n$, the ratio of joint PMF/PDF $f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta)/f_{\boldsymbol{X}^n}(\boldsymbol{y}^n, \theta)$ is constant as a function of θ (i.e. is independent of θ) if and only if $T(\boldsymbol{x}^n) = T(\boldsymbol{y}^n)$. Then $T(\boldsymbol{X}^n)$ is a minimal sufficient statistic for parameter θ.

- **Conclusion.**