ST2334 - Chapter 5 - Sampling and Sampling Distributions

5.1 - Population and Sample

5.2 - Random Sampling

5.3 - Sampling Distribution of Sample Mean

5.4 - Central Limit Theorem and its Application

5.5 - Sampling Distribution of the Difference of Two Sample Means

5.6 - Chi-square Distribution

5.7 - The Sampling Distribution of $(n-1)S^2/\sigma^2$

5.8 - The t-distribution

5.9 - The F-distribution

5.1 – Population and Sample

- Sample $\subset$ population
- Two kinds of populations : finite and infinite

5.2 - Random Sampling

- Simple random sample: every subset of n observations of the population has the same probability of being selected
- Sampling from a finite population:
  1. Sampling without replacement
     - ${}_{N}C_{n}$ samples of size n
     - Each sample has probability of $\frac{1}{{}_{N}C_{n}}$ of being selected
  2. Sampling with replacement
     - N^n samples of size n
     - Each sample has probability of $\frac{1}{N^n}$ of being selected
- Sampling from an infinite population
  - Difficult to explain random sample

5.3 - Sampling Distribution of Sample Mean

- Main purpose in selecting random samples : elicit information about the unknown population parameters
- Statistic : a function of a random sample $(X_1, X_2, \dots, X_n)$
- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- A statistic is a random variable
- Sampling distribution : probability distribution of a statistic
- Sample mean : $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- If the values in a random sample are observed and they are $x_1, x_2, \dots, x_n$, then the realization of the statistic $\bar{X}$ is given by :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Theorem 5.1** : for random samples taken from an infinite population or from a finite population with replacement having population mean μ and population standard deviation σ, the sampling distribution of the sample mean $\bar{X}$ has mean and variance : $\mu_{\bar{X}} = \mu_X$ and $\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}$

  i.e. $E(\bar{X}) = E(X)$ and $V(\bar{X}) = \frac{V(X)}{n}$

- Law of Large Number (LLN) : Let $X_1, X_2, \dots, X_n$ be a random sample of size n from a population having any distribution with mean μ and finite population variance σ^2. Then for any $\epsilon \in \mathbb{R}$,

$$P(|\bar{X} - \mu| > \epsilon) \to 0 \text{ as } n \to \infty$$

As sample size increases, probability that sample mean differs from population mean goes to zero

5.4 - Central Limit Theorem and its Application

- Central Limit Theorem : Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from a population having any distribution with mean μ and finite population variance σ^2.

  The sampling distribution of the sample mean $\bar{X}$ is approximately normal with mean μ and variance σ^2/n if n is sufficiently large.

  Hence $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ follows approximately $N(0,1)$

- Theorem 5.2 :

  ① If X_i, $i=1, 2, \cdots, n$ are $N(\mu, \sigma^2)$, then $\bar{X}$ is $N(\mu, \frac{\sigma^2}{n})$ regardless of sample size n

  ② If X_i, $i=1, 2, \cdots, n$ are approximately $N(\mu, \sigma^2)$, then $\bar{X}$ is approximately $N(\mu, \frac{\sigma^2}{n})$ regardless of sample size n

5.5 - Sampling Distribution of the Difference of Two Sample Means

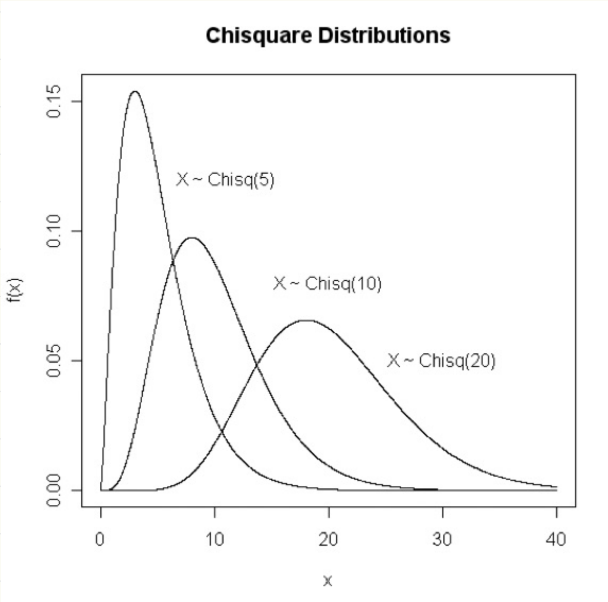
- **Theorem 5.3**: If independent samples of sizes n_1 ($\geqslant 30$) and n_2 ($\geqslant 30$) are drawn from two populations, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively, then the sampling distribution of the differences of the sample means, $\bar{X}_1$ and $\bar{X}_2$, is approximately normally distributed with mean and standard deviation given by

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 \quad \text{and} \quad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

5.6 – Chi-square Distribution

- **Definition 5.3**: If Y is a random variable with probability density function $f_Y(y) = \frac{1}{2^{n/2}\Gamma(n/2)} y^{n/2-1} e^{-y/2}$, for $y > 0$, and 0 otherwise, then Y is defined to have a chi-square distribution with n degrees of freedom, denoted by $\chi^2(n)$, where n is a positive integer, and $\Gamma(\cdot)$ is the gamma function.

- The gamma function $\Gamma(\cdot)$ is defined by
  $\Gamma(n) = \int_0^\infty x^{n-1} e^{-x}\, dx = (n-1)!$ for $n = 1, 2, 3, \ldots$



- Some properties of Chi-square distributions

① $Y \sim \chi^2(n) \longrightarrow E(Y) = n$ and $V(Y) = 2n$

② For large n, $\chi^2(n)$ approx $\sim N(n, 2n)$

③ If $Y_1, Y_2, \ldots, Y_k$ are independent chi-square random variables with $n_1, n_2, \ldots, n_k$ degrees of freedom respectively, then $Y_1 + Y_2 + \ldots + Y_k$ has a chi-square distribution with $n_1 + n_2 + \ldots + n_k$ degrees of freedom, i.e.

$$\sum_{i=1}^{k} Y_i \sim \chi^2\left(\sum_{i=1}^{k} n_i\right)$$

- Theorem 5.5 :

① $X \sim N(0,1) \longrightarrow X^2 \sim \chi^2(1)$

② Let $X \sim N(\mu, \sigma^2)$, then $[(X-\mu)/\sigma]^2 \sim \chi^2(1)$

③ Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with mean μ, and variance σ^2. Define $Y = \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2}$.

Then $Y \sim \chi^2(n)$

5.7 - The Sampling Distribution of $(n-1)S^2/\sigma^2$

- Let $X_1, X_2, \dots, X_n$ be a random sample from a population. Then the statistic $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is the sample variance.
- The sampling distribution of random variable S^2 has little practical application
- Instead, consider the sampling distribution of the random variable $\frac{(n-1)S^2}{\sigma^2}$ when $X_i \sim N(\mu, \sigma^2)$ for all i.
- **Theorem 5.5**: if S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the random variable $\frac{(n-1)S^2}{\sigma^2}$ has a chi-square distribution with $n-1$ degrees of freedom, i.e. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

5.8 - The t-distribution

- **Definition 5.4**: Suppose $Z \sim N(0,1)$ and $U \sim \chi^2(n)$. If Z and U are independent, and let $T = \frac{Z}{\sqrt{U/n}}$. Then the random variable T follows the t-distribution with n degrees of freedom, i.e. $\frac{Z}{\sqrt{U/n}} \sim t(n)$

- p.d.f. of a t-distribution: If T follows a t-distribution with n degrees of freedom, then its p.d.f. is given by

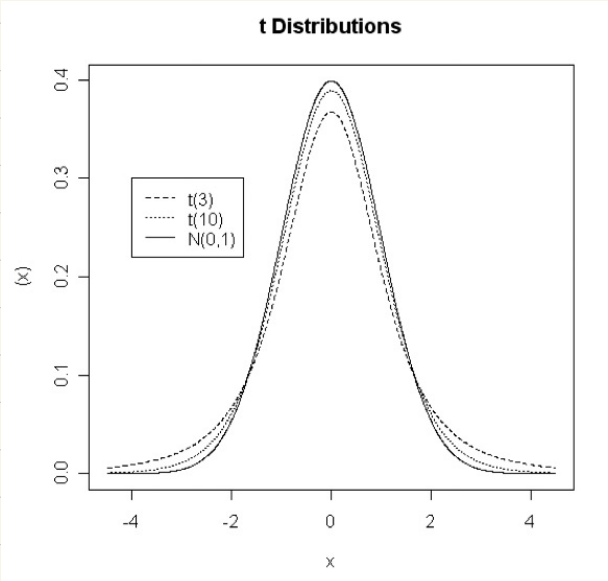
$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < t < \infty$$

The gamma function $\Gamma(\cdot)$ is defined by

$$\Gamma(n) = \int_0^\infty x^{n-1}e^{-x}\,dx = (n-1)! \quad \text{for } n = 1, 2, 3, \ldots$$

- Properties of a t-distribution

① Graph is symmetric about the vertical axis and resembles the graph of the standard normal distribution



② p.d.f. of t-distribution with n.d.f. is approaching to the p.d.f. of standard normal distribution when $n \to \infty$, i.e.

$$\lim_{n\to\infty} f_T(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \quad \text{as } n \to \infty .$$

③ The values of $Pr(T \geq t) = \int_t^{\infty} f_T(x)\,dx$ for selected values of n and t are given in a statistical table

④ $T \sim t(n) \longrightarrow E(T) = 0$ and $V(T) = n/(n-2)$ for $n > 2$

- If the random sample was selected from a normal population, then

$$Z = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} \sim N(0,1) \quad \text{and} \quad U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

It can be shown that $\bar{X}$ and S^2 are independent, and so are Z and U.

$$\therefore \; T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{\frac{(n-1)S^2}{\sigma^2} / (n-1)}}$$

$$= \frac{Z}{\sqrt{U/(n-1)}} \sim t_{n-1}$$

i.e. T has a t-distribution with $n-1$ d.f.

5.9 - The F-distribution

- **Definition 5.5**: Let U and V be independent random variables having $\chi^2(n_1)$ and $\chi^2(n_2)$ respectively, then the distribution of the random variable $F = \frac{U/n_1}{V/n_2}$, is called an F-distribution with (n_1, n_2) degrees of freedom.

  p.d.f. of F is given by

  $$f_F(x) = \frac{n_1^{n_1/2} \, n_2^{n_2/2} \, \Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \frac{x^{n_1/2-1}}{(n_1 x + n_2)^{(n_1+n_2)/2}},$$

  for $x > 0$ and 0 otherwise.

  It can be shown that $E(X) = n_2/(n_2-2)$, with $n_2 > 2$ and

  $$V(X) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}, \text{ with } n_2 > 4$$

- **Theorem 5.7**: $F \sim F(n,m) \longrightarrow 1/F \sim F(m,n)$

  This theorem follows immediately from the definition of F-distribution

  Values of the F-distribution can be found in the statistical tables

  The table gives the values of $F(n_1, n_2; \alpha)$ s.t.

  $$Pr(F > F(n_1, n_2; \alpha)) = \alpha$$

- **Theorem 5.8**: $F(n_1, n_2; 1-\alpha) = 1/F(n_2, n_1; \alpha)$