Introduction to Statistics for Data Science
Hypothesis Testing

Zehao Qian

October 26, 2023

1 Recall: Normal Distribution

1.1 Learning Materials

1. How reliable are the values found?
2. How reliable are the conclusions drawn?

1.2 Learning Objectives

1. Define the chi-squared distribution
2. Define the T-distribution
3. Define the F-distribution

1.3 Recall: The Normal Distribution

- Denote by $X \sim N(\mu, \sigma^2)$
- $E(X) = \mu$, $Var(X) = \sigma^2$
-
$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$
- Standard normal:$Z \sim N(0, 1)$
- $E(Z) = 0$, $Var(Z) = 1$
-
$$\frac{X-\mu}{\sigma} = Z, \textbf{Standardisation}$$

1.4 Chi-Squared Distribution

AKA Chi-square distribution, AKA χ^2 Distribution

- Resuqired one parameter,natural number k
- $$U_k \sim \chi_k^2$$
- Z_1, Z_2, Z_3,... ,Z_k: k independent standard normal random variables
- $$U_k = \sum_{i=1}^{k} Z_i^2$$

1.4.1 The degree of Freedom

Conception: The number of independent values available to us when calculate a value

For U_k we need k independent values,Z_1, Z_2, Z_3,... ,Z_k hense k degree of freedom

1.5 t-Distribution

AKA T distribution, AKA stident's t-Distribution

- Resuqired one parameter, degrees of freedom k
- $$T_k \sim t_k$$
- $$T_k = \frac{Z}{\sqrt{\frac{U_k}{k}}}$$

1.6 F-Distribution

- Resuqired two parameters, degrees of freedom m and n
- $$F_{m,n} \sim F_{m,n}$$
- $$F_{m,n} = \frac{U_m}{U_n}$$

2 Sample

What is the use of a sample?

- Can use to calculate statistics.
- Hope those statistics have values close to true parameter of population.

2.1 Learning Objectives

1. Get familiar with various aproaches to sampling
2. Define and make use of sampling distribution
3. Exress and make use of the **Central Limit Theorem**

2.2 Why sampling?

2.3 Types of sampling

- simple random sampling
- stratified sampling
- cluster sampling
- multistage sampling
- Non-random sampling
  - Convenience sampling
  - Judgement sampling
  - Quota sampling

2.4 The sampling distribution

Function of random variables

- If X_1, X_2,... ,X_n are random variables
- $$f(X_1, X_2, ..., X_n) = \frac{\sum_{i=1}^{n} X_i}{n} = Y$$
- Y must also be a random variables
- Y must also be the mean value, $\bar{X}$
- The mean of RVs is also an RV

The distribution of a statistic over repeated samples

- proportion $\hat{p}$
- Variance s^2
- $$E(\bar{X}) = \mu_{\bar{X}} = \mu$$

2.5 The Central Limit Theorem (CLT)

- Let X be any random variables, $E(X) = \mu$, $Var(X) = \sigma^2$
- Collect n realisation of X to get $\bar{x}$
- $$\textbf{Set}\bar{X_n} = \sum_{i-1}^{n} \frac{X_i}{n}$$

Central Limit Theorem: for a large n, the sampling distribution of $\bar{X_n}$ is approximately $N(\mu, \frac{\sigma^2}{n})$

$$lim_{n\to\infty}\bar{X_n} \sim N(\mu, \frac{\sigma^2}{n})$$

2.5.1 Three points about the CLT

1. n should be at least 30
2. a sum of normal RVs is a normal RV, A normal RV divided by n is a normal RV
   Each X_i normal $\Rightarrow$ $\bar{X}$ normal
   $\underline{X_n} \sim N(\mu, \frac{\sigma^2}{n})$ true for all value of n
3. Cab standardise a normal RV X
4. $$lim_{n\to\infty}\bar{X_n} \sim N(\mu, \frac{\sigma^2}{n}) \Rightarrow \lim_{n\to\infty} \frac{\bar{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

2.5.2 Infinite Population

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} \Leftarrow$$

Makes no difference and ALWAYS TRUE under condition above

2.5.3 Finite Population

Each value sampled changes nature of remaining unsampled popukation.

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \times \frac{N-n}{N-1}$$

- N is size of population
- Finite population
- $\sigma_{\bar{x}}^2$ for finite population with variance $\sigma^2 < \sigma_{\bar{x}}^2$ infinite population with variance σ^2
- As n increases, $\sigma_{\bar{X}}^2$ gets smaller for same N and σ

2.6 Standard Error

2.7 Variance of A Sampling Distribution

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\frac{s^2(n-1)}{\sigma^2} \sim \chi_{n-1}^2$$

$$E(s^2) = \frac{\sigma^2}{n-1}E(\chi_{n-1}^2)$$

Due to the fact that $E(\chi_k^2) = k$, so $E(s^2) = \sigma^2$

3 Estimation

3.1 Types of Estimation

- Point Estimation
- Interval Estimation

3.2 Learning Objects

1. Explore