

# W1211 Introduction to Statistics

## Lecture 17

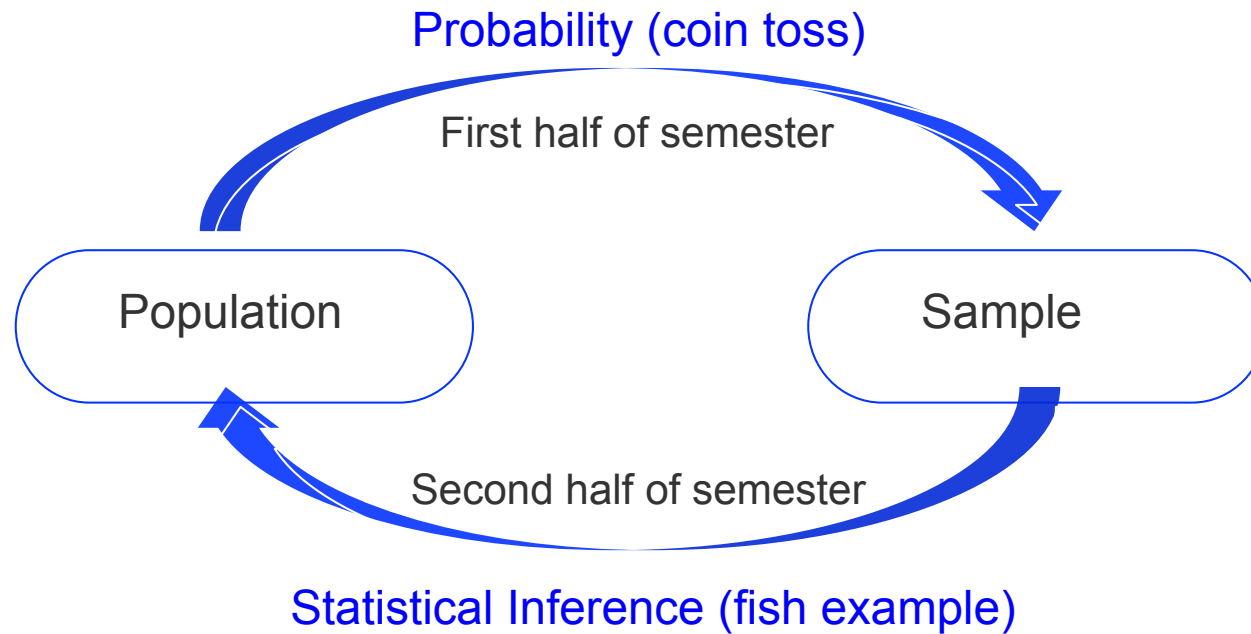
Wei Wang

Oct 31st, 2012

# Population and Sample

- ▶ We will start changing our discussion from probability to statistics, which means we need to think about samples and how they relate to the underlying population.
- ▶ Recall the relationship between population and sample (probability and inference) that we visualized in the first lecture.

# Probability and Inference



# RV or a Particular Number

- ▶ In the first chapter, we use lowercase letters to represent the sample,  $x_1, x_2, x_3, \dots$ . That means we have already observed the data and each of the letters can be replaced by a particular number.
- ▶ Before the data becoming available, there is uncertainty as to what value we will observe, so we view each observation as a RV, thus denoted by uppercase letter  $X_1, X_2, X_3, \dots$ .

# Sample and Statistics

- ▶ A statistic is any quantity whose value can be calculated from sample data, such as Sample Mean and Sample Variance.
- ▶ Before obtaining data, a statistic is also a RV. The bulk of statistical inference is to find the distribution of the statistics, or the so-called *Sampling Distributions*.
- ▶ To make things easier, we often need to assume the observed data are *Simple Random Samples*, which means they are IID (Independently Identically Distributed).

# Introduction to IID

- A sequence of random variables,  $X_1, X_2, \dots, X_n$ , is **independent and identically distributed (i.i.d.)** if each random variable has the same probability distribution as the others and all are **mutually independent**.
- In statistical analysis, we often assume the sampled data  $X_1, X_2, \dots, X_n$ , are i.i.d. from a common distribution  $f(x)$ . And usually, we end up analyzing a **linear combination** of the  $X_i$ 's, that is

$$Y = a_1X_1 + \dots + a_nX_n = \sum_{i=1}^n a_iX_i$$

# Deriving a Sampling Distribution

- ▶ Probability rules can be used to obtain the distribution of a statistic that is a fairly simple function of  $X_i$ 's and the distribution of  $X_i$ 's are also of simple form. See Example 5.20 on page 215.
- ▶ However, this brute force method only works for a limited class of statistics. For most statistics, it is practically impossible to get the distribution (pmf/pdf).
- ▶ But for some of the most common statistics, such as Sample Mean, we have some useful results.

# Sample Mean

- ▶ Let  $X_1, X_2, \dots, X_n$  be an IID sequence of random variables from a distribution with mean  $\mu$  and variance  $\sigma^2$ . Sample Mean is

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}.$$

- ▶ Then we can derive the expectation and variance of sample mean  $\bar{X}$

$$E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \mu$$

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$



# Invariance of Normal RV under Linear Transformation

- ▶ When  $X_1, X_2, X_3, X_4, \dots$  are normal random variables, then the linear combination of them

$$a_1X_1 + a_2X_2 + \dots + a_nX_n = \sum_{i=1}^n a_iX_i$$

is still a normal random variable.

- ▶ In particular, sample mean  $\bar{X}$  is still a random variables.
- ▶ Remark: 1. No IID assumption is necessary; 2. This property is for Normal only.

# Sample Mean of IID Normal

- ▶ If  $X_1, X_2, \dots, X_n$  IID  $\sim N(\mu, \sigma^2)$ , then what is the distribution of  $\bar{X}$ ?

# Sample Mean of IID Normal

- ▶ If  $X_1, X_2, \dots, X_n$  IID  $\sim N(\mu, \sigma^2)$ , then what is the distribution of  $\bar{X}$ ?

▶

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

# Sample Mean of IID Normal

- ▶ If  $X_1, X_2, \dots, X_n$  IID  $\sim N(\mu, \sigma^2)$ , then what is the distribution of  $\bar{X}$ ?

▶

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- ▶ But how can we derive the distribution of Sample Mean when the sample are not normal? We need to use Central Limit Theorem.

# CLT

- Theorem:

The Central Limit Theorem (CLT)

Let  $X_1, X_2, \dots, X_n$ , be an i.i.d. sequence from a distribution with mean  $\mu$  and variance  $\sigma^2$ . Then if  $n$  is sufficiently large, the sample mean  $\bar{X}$  has approximately a normal distribution with  $\mu_{\bar{X}} = \mu$  and  $\sigma_{\bar{X}}^2 = \sigma^2/n$ ; And the sample total has approximately a normal distribution with  $\mu_T = n\mu$ ,  $\sigma_T^2 = n\sigma^2$ . The larger the value of  $n$ , the better the approximation.

- Rule of Thumb: if  $n > 30$ , the CLT can be used.

# Distribution of a Linear Combination

- ▶ Sample mean is a particular case of linear combinations.
- ▶ The expectation and variance of a general linear combination

$$a_1X_1 + a_2X_2 + \dots + a_nX_n$$

is given by the following result.

## A key result \*\*\*

Let  $X_1, X_2, \dots, X_n$ , have mean values  $\mu_1, \mu_2, \dots, \mu_n$ , respectively, and variances  $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ , respectively.

- Whether or not the  $X_i$ 's are independent,

$$\begin{aligned} E(a_1X_1 + a_2X_2 + \dots + a_nX_n) &= a_1E(X_1) + a_2E(X_2) + \dots + a_nE(X_n) \\ &= a_1\mu_1 + a_2\mu_2 + \dots + a_n\mu_n \end{aligned}$$

- For any  $X_1, X_2, \dots, X_n$ ,

$$\text{Var}(a_1X_1 + a_2X_2 + \dots + a_nX_n) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \text{Cov}(X_i, X_j)$$

If they are independent, then

$$\begin{aligned} &\text{Var}(a_1X_1 + a_2X_2 + \dots + a_nX_n) \\ &= a_1^2 \text{Var}(X_1) + a_2^2 \text{Var}(X_2) + \dots + a_n^2 \text{Var}(X_n) \\ &= a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \dots + a_n^2 \sigma_n^2 \end{aligned}$$

# Special Cases

- $E(X+Y) = E(X) + E(Y)$ ;
- $E(X-Y) = E(X) - E(Y)$ ;
- $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$
- $\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$
- If  $X$  and  $Y$  are independent, then  $\text{Cov}(X, Y) = 0$ , and  
 $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$   
 $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$



# Example

Ex. Show that if  $X \sim \text{Bin}(n, p)$ , then  $E(X) = np$ , and  $\text{Var}(X) = np(1 - p)$ .

Ex. Show that if  $X$  is a negative binomial rv with pmf  $nb(x; r, p)$ , then  $E(X) = r(1-p)/p$ ,  
 $\text{Var}(X) = r(1 - p)/p^2$ .